Visualisation des données et prétraitements

Une image contenant diagramme, ligne, Plan, Dessin technique

Description générée automatiquement

Après les valeurs aberrantes enlevé :

Une image contenant texte, capture d’écran, ligne, Police

Description générée automatiquement

Une image contenant diagramme, ligne

Description générée automatiquement

Ensuite on enlève les points ou le moteur est à l’arrêt, et on enlève 30 points après chaque démarrage.

Une image contenant capture d’écran, texte, ligne, diagramme

Description générée automatiquement

Une image contenant diagramme, ligne

Description générée automatiquement

Puis on divise les jeux de données :

Ceux avec le mode dégradation = 1 et ceux avec le mode dégradation = 0.

On prend pour chaque jeu de donnée, 70% de valeur d’entrainement et 30% de valeur pour tester les modèles.

Analyse Régression Linéaire.

Etudions les corrélations des données : Ici on montre l’exemple avec le jeu de donnée ayant le mode dégradation = 0, mais les démarches sont similaires pour l’autre jeu de donnée.

On décide de choisir comme indicateur de santé la puissance du Moteur.

Une image contenant texte, Police, conception

Description générée automatiquement

Une image contenant texte, capture d’écran, Police, nombre

Description générée automatiquement

Une image contenant texte, croquis, diagramme, dessin

Description générée automatiquement

Matrice de corrélation des données. On observe clairement certaines relations linéaires. Par exemple x2 et y, x2 et x3.

Cependant, notre objectif est de comparé le comportement du moteur (soit l’indicateur de santé choisi) en fonction des paramètres en entrés. Etudions plus précisément les corrélations possibles.

y = train\_data\_deg0$PW

x1 = train\_data\_deg0$Mo

x2 = train\_data\_deg0$CO

x3 = train\_data\_deg0$CR

x4 = train\_data\_deg0$T1

Une image contenant texte, diagramme, croquis

Description générée automatiquement

Lorsque l’on étudie la relation linéaire multiple de chaque variable sur y, on obtient ce graphique qui permet de savoir quelles variables à une influence significative sur y.

Une image contenant Rectangle, diagramme, capture d’écran, carré

Description générée automatiquementUne image contenant texte, capture d’écran, diagramme, Rectangle

Description générée automatiquementUne image contenant diagramme, Rectangle, capture d’écran, carré

Description générée automatiquement

On observe que les variables d’entrée ont principalement toutes une influence significative sur y. Cependant on décide de s’intéresser à une régression linéaire simple dans un premier cas. Car s’il s’avère qu’une régression linéaire simple est suffisante à prédire avec précision le paramètre de santé, alors choisirons cette régression simple comme modèle, car ce type de modèle est simple et facile à comprendre.

On étudie maintenant CO pour prédire PW :

y = train\_data\_deg0$PW

x2 = train\_data\_deg0$CO

Une image contenant texte, capture d’écran, Police, document

Description générée automatiquement

On obtient une régression linéaire significative (p-value très faible) avec un R²ajusté de 0.9951, soit un modèle très bon.

Une image contenant texte, ligne, diagramme, Tracé

Description générée automatiquement

Aucun doute sur la linéarité du modèle. Voici les intervalles de confiances à 95% des paramètres estimés :

Une image contenant texte, Police, blanc, ligne

Description générée automatiquement

Maintenant on veut s’assurer que le modèle soit utilisable dans le futur. Nous allons tester sa viabilité sur le jeu de donnée de test (toujours en mode dégradation =0). Pour savoir si une valeur de PW étudié n’est pas en adéquation avec le modèle, donc qu’il y a dégradation, il faut définir des intervalles de contrôles. En se basant sur la méthodologie Lean 6 Sigma, nous définissons des limites de contrôles à + ou – 3 sigma de la variation naturelle du processus.

La variation naturelle du processus est de :



Notons que nous devrions vérifier des hypothèses (ici admise) afin de s’assurer que le modèle est viable réellement (étude de la normalité des résidus et de la non-autocorrélation).

Voici les données représentées avec les limites de contrôles. Si un point est au-dessus ou en dessous de ces limites, alors on peut considérer qu’il y a une dégradation du moteur.

Une image contenant texte, ligne, diagramme, capture d’écran

Description générée automatiquement

Pour information, 99.77% des valeurs d’entrainements sont dans l'intervalle de contrôle. Soit 25 points sont en dehors, on peut dire qu’il y a eu 25 fois un comportement non prévu, soit une dégradation du moteur sur les 70% premiers points. Vérifions maintenant sur des données que le modèle n’a jamais vues :

On obtient 99.73% des valeurs dans l’intervalles de contrôle (soit 13 points en dehors des limites de contrôles), soit une précision très élevée. L’étude de la puissance du moteur en sortie en fonction du couple voulu à l’entrée permet clairement de prédire une dégradation du moteur.

Nous savons que ce modèle fonctionne, cependant, nous devons confirmer les hypothèses de normalités des résidus et de non-autocorrélation des résidus.

Une image contenant capture d’écran, texte, diagramme, Tracé

Description générée automatiquement

Visuellement, les résidus semblent normaux. Cependant, un test de Lilliefors nous indique que les résidus ne sont absolument pas normaux (p valeur très faible), ce qui est confirmé par un qq-plot.

Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement

Une image contenant texte, ligne, diagramme, capture d’écran

Description générée automatiquement

Nous ne pouvons pas conclure à la normalité des résidus. Quand est-il de l’autocorrélation ?

Une image contenant texte, capture d’écran, diagramme, nombre

Description générée automatiquement

Une image contenant texte, Police, capture d’écran, blanc

Description générée automatiquement Une image contenant texte, Police, capture d’écran, ligne

Description générée automatiquement

Le graphique ACF et les deux tests de Box-Pierce et Box-Ljung indique que les résidus ne sont pas autocorrélé (p-valeur élevé).

Nous pouvons conclure que ce modèle est utilisable, mais qu’il faut rester prudent car si nous n’avons pas la normalité des résidus, alors il se peut qu’un autre facteur non mesuré ici puisse impacter le comportement du modèle et devenir faux à l’avenir.

Nous avons aussi essayé de faire des régressions multiples, les résultats n’étaient pas significativement meilleures, et les hypothèses d’autocorrélation et normalité nétaient pas validées, alors nous préférons garder ce modèle de régression simple.