UNIVERSITÉ PARIS SACLAY

AVRIL 2023

RAPPORT PROJET D'IAS

Prédire la satisfaction d'un passager aprés un vol d'avion

PRÉSENTÉ À

Solal Nathan

PRÉSENTÉ PAR

Sarah Boundaoui

Ahmed Baaroun

• ***INTRODUCTION:***

Ce projet est un projet à but ouvert qui consiste à choisir un Dataset libre sur lequel nous allons devoir choisir un objectif, raffiné le data approfondir la maitrise et découvrir des nouveaux algorithmes et accomplir l'objectif via un ou plusieurs modèles vu en cours.

• ***PRESENTATION DU DATASET:***

Nous avons choisi un Dataset qui contient les informations d'un survey donnant le niveau de satisfaction de passager après un vol d'avion. Le Dataset comporte environ 103 000 lignes et 24 labels, il comporte 2 fichiers.cvs, l'un pour l'entrainement des données et l'autre pour le test de donnée.

Lien vers le Dataset : https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?select=test.csv

• ***OBJECTIF* DU PROJET:**

L'objectif de notre projet est d'approfondir nos connaissances acquises au cours d'introduction à l'apprentissage statistique et aussi de découvrir de nouveaux algorithmes.
Avec ce type de donnée on peut prédire la satisfaction d''un passager après un vol d'avion et ça en suivant plusieurs étapes: charger les données et les explorer, pre-processing, Modèles de classification, optimisation, visualisation.

• ***ETAPES DU PROJET:***

1. Chargement des données:

Notre Dataset contient 103904 lignes et 24 colonnes. Parmi les 24 colonnes on a remarqué que la colonne 'id' ne sert à Rien on l'a donc enlevé, on a ensuite affiché la liste de l'ensemble des colonnes avec le type de données et les valeurs non nulles.

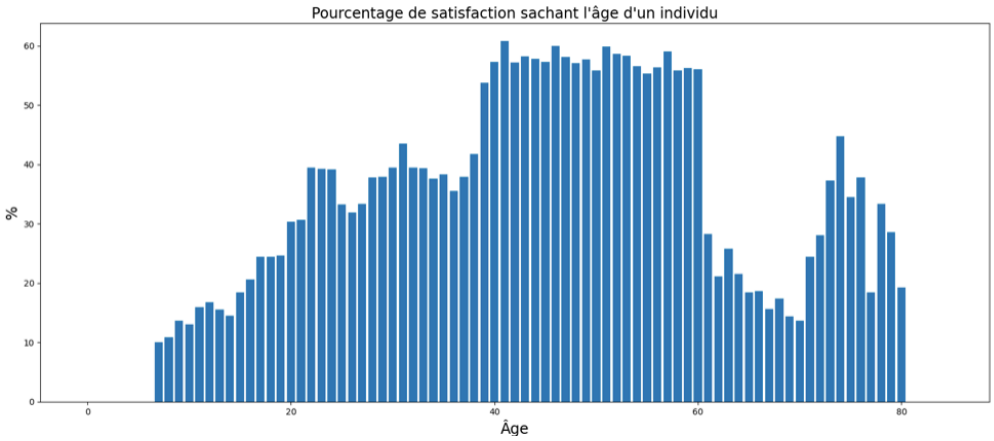
Au sein de notre jeu de données, nous remarquons que la plupart des colonnes sont catégoriques, seules quatre colonnes peuvent prendre n'importe quelles valeurs : "Age", "Flight Distance", "Departure Delay in Minutes", "Arrival Delay in Minutes ».

Nous avons également remarqué qu'il nous manque des données au sein de la colonne "Arrival in Delay Minutes" (qui contient 103 594 valeurs non nulles tandis que le reste des colonnes en ont 103 904), Nous les avons donc les retirer. Nous avons ensuite affiché les statistiques, celle de la variable "Departure Delay in Minutes" semblent démontrer que cette variable possède un grand nombre de valeurs aberrantes :
Les quartiles démontrent que la majorité des valeurs de cette variable est concentrée autour de 0 min et une petite majorité autour de 13 min. La moyenne n'est que de 15 min tandis que la valeur maximale est de 1 592 min.

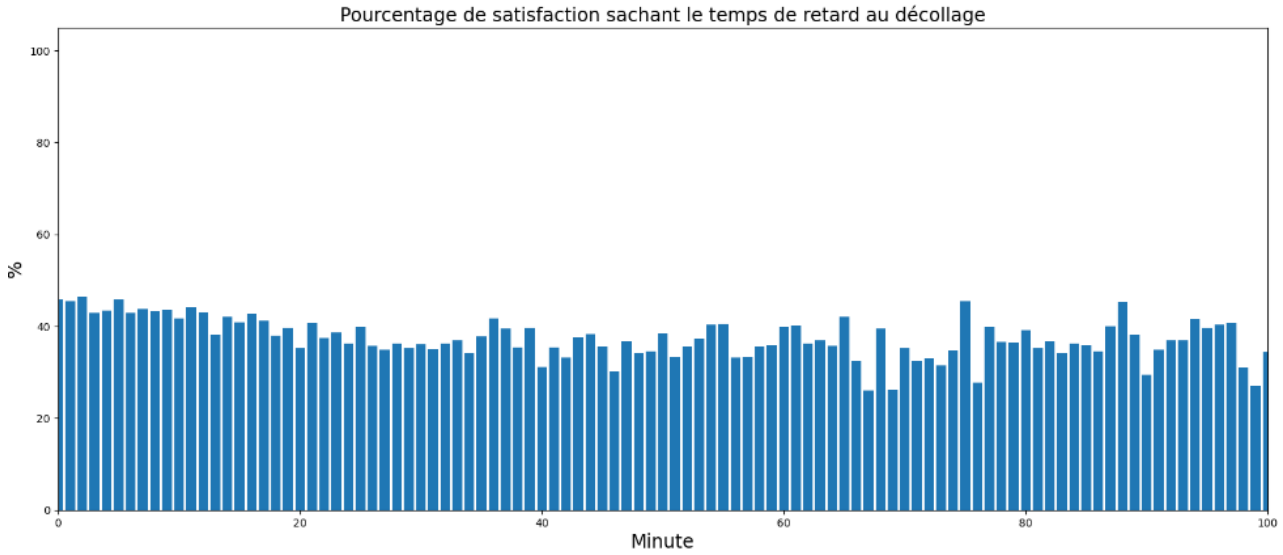
2. Exploration des données:

On remarque que sensiblement, les colonnes "Age", "Flight Distance", "Departure Delay in Minutes" et "Arrival Delay in Minutes" ont une logique de notation irrégulière, ce qui est évidemment naturel, contrairement aux autres colonnes, purement catégoriques.

Aussi, on remarque 4 colonnes dont les catégories sont nominales, soit "Gender", "Customer Type", "Type of travel" et "Class". Ainsi, on va logiquement séparer notre visualisation des données sous forme de diagrammes, en 3 parties correspondant aux différences des colonnes, voici deux exemple avec explication:

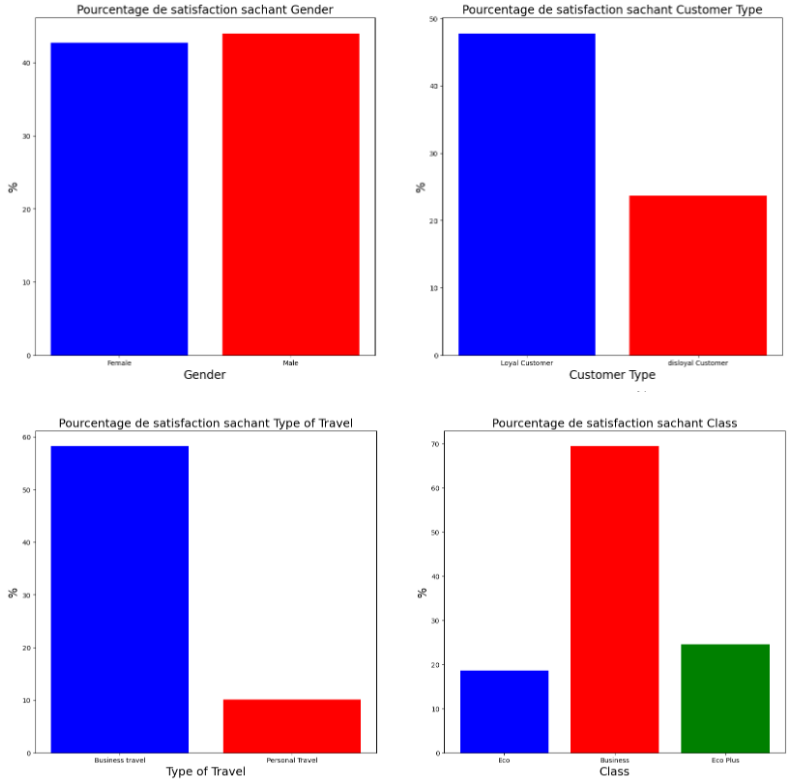


- La tranche d'individus satisfaits se trouve entre 40 et 60 ans. La probabilité y est constante et vaut près de 60%.
- De façon générale, les jeunes n'apprécient pas du tout leur vol, mais cela s'améliore avec l'âge tant que celui-ci reste plus petit que 40 ans. Le phénomène inverse se produit après 60 ans



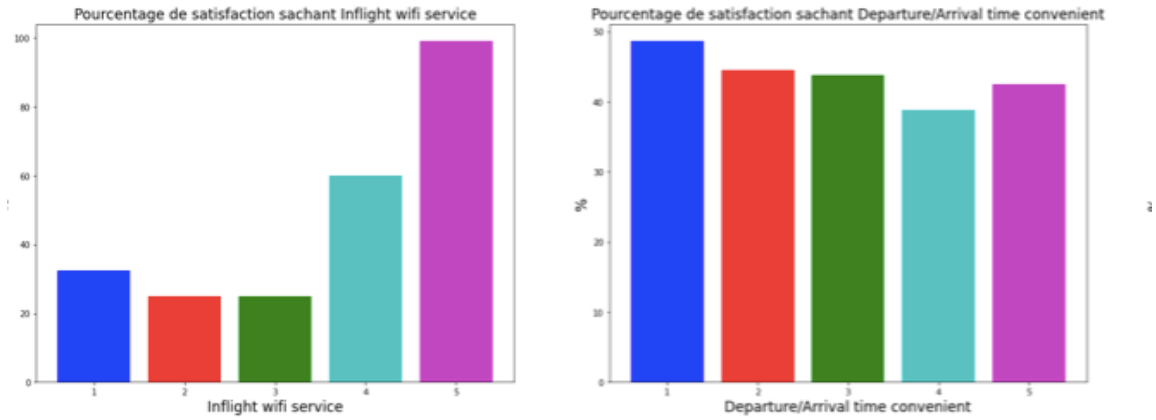
- La probabilité de satisfaction d'un individu sachant le temps de retard au décollage est très légèrement décroissante entre 0 et 40min où elle varie entre 45% et 40%. Après 40min, il est difficile de procéder à une analyse en raison des fluctuations. Elles sont dues à un manque d'individus ayant attendu un nombre de minutes données.
- On peut remarquer que la moyenne de ces probabilités a l'air de tourner autour de 40%.
- Ces observations confortent l'idée que la variable de retard au décollage ne joue pas un rôle prépondérant dans l'évaluation de la satisfaction.

Ensuite on a récupérer le pourcentage de satisfaction pour quelque colonnes : Gender', 'Customer Type', 'Type of Travel', ‘Class' , le code utilisé est bien expliqué sur le Jupyter, voici les diagrammes :



- Un habitué de la ligne de vol (« Loyal Customer") a 2 fois plus satisfait qu'un client infidèle.
- Un individu ayant effectué un voyage pour raison personnelle a 6 fois moins de chance d'être satisfait qu'un individu ayant effectué un voyage d'affaire.
- On observe également que la classe affaire garantie presque la satisfaction d'un individu (~70%). Il n'y a pas de différences trop notables entre les classes Eco et Eco Plus.
- Il n'y a pas de différence notable entre Féminin et Masculin pour la satisfaction.

On a aussi récupérer le pourcentage pour le reste des données : on a eu 14 figures, voici deux exemples :



- En général, les clients sont plus satisfaits lorsque les différentes caractéristiques du service obtiennent des notes élevées, sauf pour "Gate Location" et "Departure/Arrival Time convenient". En particulier, une note élevée pour "Inflight wifi service" est pratiquement une garantie de satisfaction à 99% pour un individu.

- On observe une tendance similaire, bien que moins marquée, avec "Ease of Online Booking" et "Online boarding". En revanche, une mauvaise note pour "Online boarding" et "Inflight entertainment" conduit très probablement à l'insatisfaction d'un individu.

- En ce qui concerne "Departure/Arrival Time convenient", les probabilités constantes autour de 45% suggèrent que cette caractéristique ne joue pas un rôle majeur dans l'explication de la satisfaction.

On passe ensuite à la corrélation d'abord on transforme les données de satisfaction en int, après on crée la heatmap:

Nous constatons que la satisfaction des clients n'est pas corrélée avec les caractéristiques telles que "Departure/Arrival time convenient", "Gate location", "Departure Delay in Minutes", et "Arrival Delay in Minutes".

Par conséquent, nous allons les supprimer de notre analyse. Cependant, il existe toujours une forte corrélation entre la satisfaction et de nombreuses autres caractéristiques. Il est important de noter que la corrélation entre "Departure Delay in Minutes" et "Arrival Delay in Minutes" devrait normalement être de 1.

Cependant, sur le tableau de corrélation, nous observons une valeur de 0,97, qui est probablement due à des erreurs d'arrondi. Il est également possible que les corrélations entre les caractéristiques ne soient pas linéaires, mais pour faciliter l'analyse, nous supposerons qu'elles le sont.

3. Pre-processsing:

Transformation de la colonne satisfaction en int, Creation des train/test set, validation set (80-20), Au final on a 82875 lignes et 18 colonnes

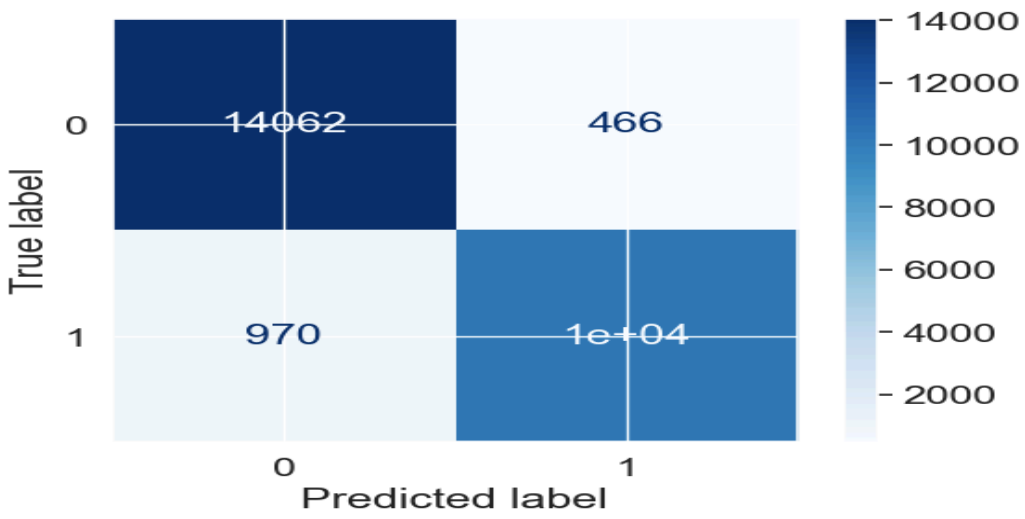
4. Model de classification:

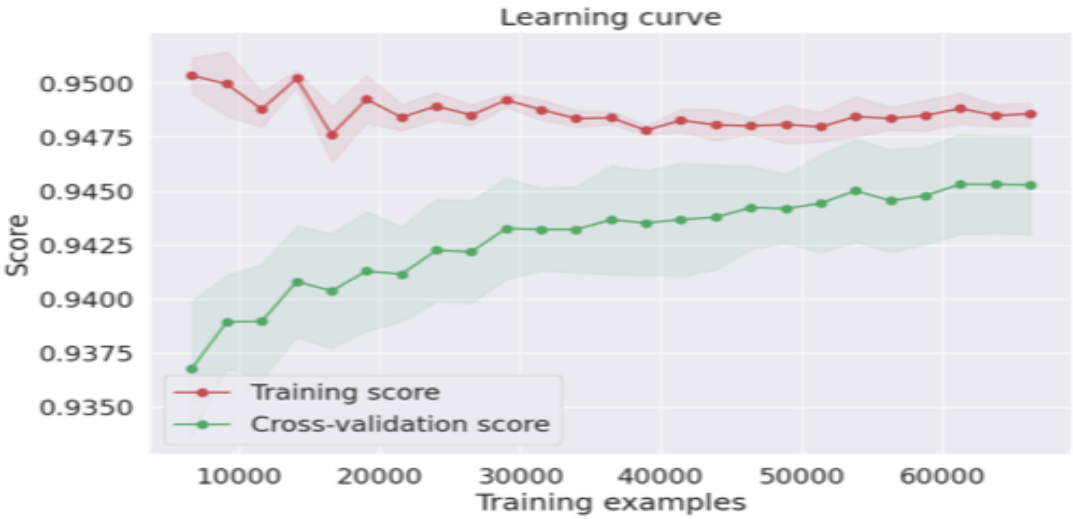
Maintenant on passe à la classification, il existe plusieurs méthode en machine learning pour prédire la satisfaction d'un passager après un vol d'avion, Sachant la taille importante de notre dataset, nous avons décidé d'appliquer deux modèles d'algorithmes d'apprentissage différents à savoir "Decision Tree" en tant que Classifieur, étant donné la "binarité" du résultat (satisfaction ou pas), et "Random Forest", qui n'est ni plus ni moins que la concaténation de plusieurs "Decision Trees".

Arbre de décision : c'est un model qui peut être utilisé pour les problèmes de classification avec plusieurs classes, Cette méthode est facile à comprendre et à interpréter, ce qui peut aider les analystes à identifier les caractéristiques les plus importantes pour prédire la satisfaction des passagers. Cette méthode permet de classer les données en utilisant un arbre où chaque nœud représente une décision et chaque feuille représente une classe.

Aprés l'application du modèle on a une précision de 90%, ce qui constitue déjà un score non-négligeable. Or nous allons essayer d'optimiser ce résultat en mettant en place des hyperparamètres plus "affinés" en utilisant la validation croisée et élilminer le plus de faux positifs et de faux négatifs tout en se concentrant en priorité sur les faux positifs, qu'on cherche à éliminer au mieux.
Ensuite avec les nouveaux hyperparamètres après optimisation on est passé à une précision de 94%

On a ensuite affiché la matrice, et on affiche les train score et validation score.



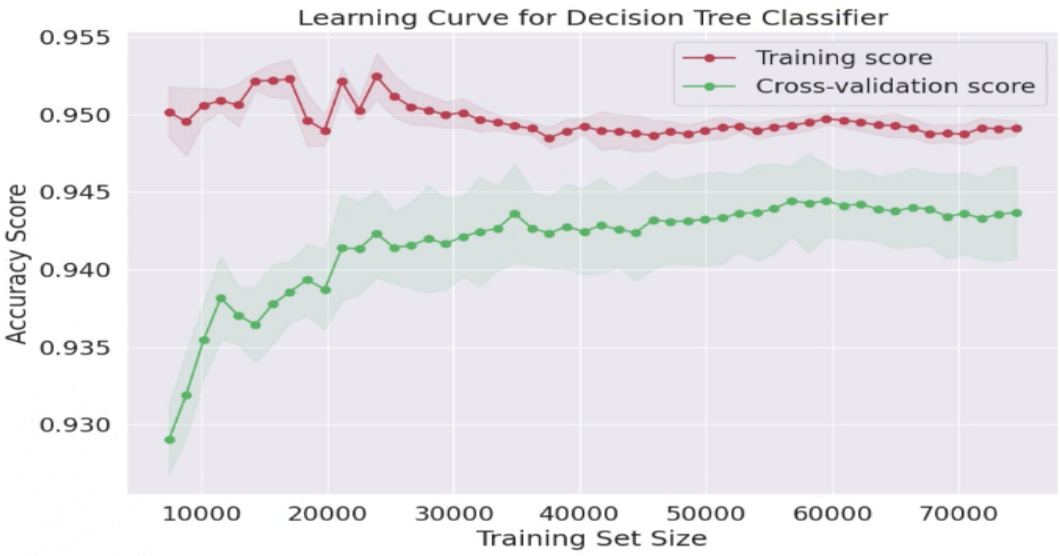


métrique de score est l'accuracy.

- Validation score a l'air de continuer à augmenter, peut-être qu'avec plus de données on obtiendra un meilleur score
- avec la méthode Decision tree on a un pourcentage de 98 %

Random forest : les forêts aléatoires sont une extension des arbres de décision et elles sont souvent utilisées pour des problèmes de classification avec de nombreuses classes. Cette méthode peut aider à réduire le surajustement et peut être utilisée pour classer les caractéristiques les plus importantes.

Juste après l'application de la méthode randome forest on a un pourcentage de 94% un score qui est très bien, que même avec une optimisation on ne peut avoir mieux. On a ensuite affiché train et validation score



Le train score est presque stable.

- ***CONCLUSION :***

Pour finir on a fait une comparaison entre les deux méthodes pour savoir laquelle des deux est meilleures, Avant même de regarder le résultat de ROC AUC on peut dire que le random forest est meilleur que decision tree, car dés la première application on a eu 94% mais avec decision tree on a eu 90% c'est aprés optimisation qu'on est passée à 94%. Même après l'affichage de la courbe ROC on voit que Random forest reste meilleur avec un pourcentage de 98,9%, mieux que Decision tree avec un pourcentage de 98,6%, avec plus de données on peut avoir un plus grand écart entre les deux modèles.