UFR 6
Université Paul Valéry, Montpellier III

Rapport de projet

Kaggle Competition : Spaceship Titanic

Jonathan Duckes

January 2024

La réalisation de ce projet s'inspire très fortement de la réalisation faite par les utilisateurs Kaggle 'Gusthema' et ''eliot robot' pour la compétition Kaggle [1]"Spaceship Titanic".

Remerciements

Je tiens tout d'abord à remercier les responsables de la formation Madame Sophie Lèbre, Monsieur Maximilien Servajean et Monsieur Arnaud Sallabery pour leur encadrement, leur disponibilité et surtout pour m'avoir donnée la possibilité de travailler sur un projet compte tenu de ma situation.
Je souhaite aussi exprimer ma gratitude envers mes collègues et camarades de classe, m'ayant fourni des conseils et jugements intéressants afin d'avancer dans ce projet.
Mes remerciements sont également dirigés vers l'équipe pédagogique et au corps enseignant de la formation, pour leur enseignement et la transmission des outils et connaissances qui ont été d'une grande utilité pour mener à bien ce projet.

Résumé

L'alternance en Master MIASHS offrent aux étudiants une expérience pratique dans l'application des mathématiques et de l'informatique aux sciences humaines et sociales. L'objectif est de développer davantage les compétences professionnelles des étudiants et de leur permettre de mettre en pratique leurs connaissances théoriques. Les alternances sont généralement axés sur l'exploration et l'analyse de données, la modélisation mathématique ou l'utilisation d'outils informatiques pour résoudre des problèmes du monde réel. Cela permet aux étudiants de se familiariser avec le monde du travail et de participer à des projets de recherche en cours. Ils sont encadrés par un tuteur qui les accompagne tout au long du stage.

Malheureusement pour mon premier semestre en première année de Master MIASHS je n'ai pas pu trouver d'alternance, cependant, grâce à l'aide précieuse de mes professeurs, j'ai pu trouver une alternative constructive. Ils m'ont généreusement proposé de travailler sur un projet précis. En effet ils m'ont invité à trouver un projet Kaggle, qui est une plateforme dédiée aux data scientists et à l'apprentissage automatique. La plateforme permet de participer à des compétitions de Machine Learning et de réaliser des projets. De ce fait Kaggle est très utile afin d'appliquer des connaissances en apprentissage automatique.

Afin de répondre aux exigences de mes professeurs, je vais indiquer dans ce mémoire les visualisations et statistiques descriptives que j'ai mises en place pour comprendre et analyser le jeu de données. Ce mémoire va aussi contenir des descriptions des solutions statistiques et de machine learning réalisées afin d'obtenir un modèle prédictif tout en contenant aussi des descriptions, des métriques, les tâches, le contexte, etc.

Table des matières

Table des figures

Liste des tableaux

Introduction

Afin de répondre à l'importance du choix d'un projet Kaggle pour pratiquer l'aprentissage automatique et l'analyse de données, j'ai choisi le projet "Spaceship Titanic" car non seulement il est captivant, mais il est également pertinent dans le contexte de l'apprentissage Ce projet propose une énigme fictive, dans laquelle des passagers du vaisseau spatial ont été téléportés vers une dimension parallèle à la suite d'une collision cosmique. Cette situation unique représente un défi pour les compétences en science des données, combinant à la fois des éléments de classification et de compréhension complexe de modèles. La problématique étant la nécessité de prédire avec précision quel passager a été transporté dans cette dimension parallèle souligne des enjeux de résolution de problèmes complexes, ce qui en fait un choix idéal pour l'application des techniques avancées d'apprentissage automatique. Ce projet m'a permis d'appliquer et de renforcer des connaissances en apprentissage automatique acquises au cours de la formation. En travaillant sur des données issues d'une situation cosmique fictive, j'ai été confronté à des missions réalistes qui exigent une combinaison de compétences en prétraitement des données, en modélisation et en évaluation des performances. De plus la nécessité de soumettre les prédictions directement sur Kaggle renforce l'aspect pratique de l'apprentissage, et permet ainsi de traduire la compréhension du problème en solutions pertinantes et à évaluer la performance par rapport à d'autres compétiteurs du challenge.

L'ensemble de données détaillé comprend des informations sur les passagers du Spaceship Titanic avant l'incident. Ces données comprennent des détails sur le départ des passagers depuis leur planète d'origine, leur choix de sommeil cryogénique, les installations de luxe auxquelles ils ont accès, etc. Afin d'analyser ces données, l'utilisation de TensorFlow Decision Forests, une bibliothèque puissante pour la modélisation et l'analyse prédictive, a été recommandé. Cette library permet de créer un modèle robuste pouvant prédire quels passagers ont été transportés dans cette dimension parallèle.

Afin de répondre à la problématique, l'exploration des relations complexes entre les différentes caractéristiques des passagers pour construire un modèle prédictif fiable est nécessaire. Cette tâche nécessite une compréhension approfondie des données et une mise en œuvre efficace des techniques avancées d'apprentissage automatique.

Le mémoire contiendra donc une synthèse des travaux pertinents pour le projet, la méthodologie utilisé lors de l'aboutissement de ce projet, l'analyse exploratoire des données, la construction et la sélection du modèle, les résultats et une discussion de ceux-ci.

Chapitre 1

Méthodologie

Sommaire

1.1 Méthodologie

Étant donné qu'il s'agit d'une compétition Kaggle, j'ai dû créer un compte afin de pouvoir participer à la compétition. Suite à cela, les données ,en format CSV, ont été rendues accessibles sur la page de la compétition où l'on peut trouver le dossier ZIP contenant les fichiers, une mise en contexte fictive du projet, une explication de ce qui est attendu et aussi le classement des participants. Le classement nous permet de récupérer un Notebook, pour bien débuter (To get started) [2], étant celui du participant qui a récolté le meilleur score sur ce projet. Ce Notebook contient des bouts de code, qui quand exécutés nous permettent de charger le jeu de donnée, de faire une analyse exploratoire brève, de procéder à une base de prétraitement des données, de configurer un modèle, de l'entraîner, de le tester et enfin de générer les prédictions dans un fichier CSV afin de les soumettre sur Kaggle.

Le Notebook étant pratiquement complet, j'ai cherché à rajouter des bouts de codes afin de collecter plus d'informations sur le jeu de données, ce qui me permettraient de trouver ce qui pourrait améliorer le modèle afin de le rendre plus précis ou plus stable. Pour ce fait je me suis inspiré de mes cours de la formation MIASHS et du TER afin de voir ce qui n'a pas été fait et que je pourrais donc introduire dans le Notebook tout en restant pertinent pour la réalisation du projet. J'ai donc téléchargé le dossier ZIP, qui contient les fichiers train.csv et test.csv, que je vais présenter dans la prochaine partie et le Notebook du projet. Ce dernier, contenait plusieurs outils que je devais importer tels, :

— Pandas qui permet d'analyser des données [3],
— Numpy qui permet de manipuler des tableaux et des matrices [4],
— Seaborn [5] et Matplotlib [6] qui permettent de créer des visualisations.

Or j'ai rencontré un problème avec les outils Tensorflow et Tensorflow Decision Forest, respectivement, un outil de création de modèle de ML et une librairie permettant d'entraîner des arbres de décision, mais dont le dernier n'est pas disponible sur Windows. J'ai donc procéder à l'installation de WSL, un sous système Windows sur Linux, afin de d'installer TF-DF or pour ce faire il a fallu suivre différentes étapes qui prenaient énormément de temps et aussi apprendre le système Linux. J'ai donc opté pour la solution la plus rapide qui est d'utiliser Google Colab de par son environnement qui est en ligne et non local dont le seul point négatif a été de réinstaller TF et TF-DF à chaque réouverture du Notebook.

Afin de mener à bien ce projet j'ai donc utilisé le Notebook fourni par Kaggle et contenant déjà un modèle définit et performant avec une bonne Accuracy. J'ai quand même cherché à mieux comprendre les données, à les manipuler, à observer ce qui en ressortait et ce qui pourrait être pertinent pour le modèle, que j'ai cherché à comprendre afin de pouvoir l'améliorer en lui donnant du matériel cohérent pour son entraînement. J'ai fait mes recherches à partir du forum de discussions [7] et du chat du serveur Discord de Kaggle [8] liés à la compétition. Je me suis aussi servi de StackOverFlow [9] et ChatGPT

[10] pour résoudre des erreurs et pour débugguer des bouts de codes, mais aussi pour l'interprétation de résultats. Mes cours de Master MIASHS m'ont aussi beaucoup aidé pour l'interprétation des résultats.

Dans le prochain chapitre sera presenté une exploration exhaustive des données.

Chapitre 2

Analyse Exploratoire des Données

Sommaire

2.1 Les Données

Présentation des colonnes, de leur signification, et des types de Données :

— PassengerId : Identifiant unique pour chaque passager. Entier
— HomePlanet : Planète d'origine du passager. Chaîne de caractère
— CryoSleep : Indique si le passager a été en sommeil cryogénique. Booléen
— Cabin : Détails de la cabine, comprenant Deck/Cabin_num/Side. Chaîne de caractère
— Destination : Destination finale du voyage. Chaîne de caractère
— Age : Âge du passager. Entier
— VIP : Indique si le passager a un statut VIP. Booléen
— RoomService : Nombre de fois où le service en chambre a été utilisé. Entier
— FoodCourt : Nombre de visites au food court. Entier
— ShoppingMall : Nombre de visites au centre commercial.
— Spa : Nombre de visites au spa. Entier
— VRDeck : Nombre d'utilisations du pont VR. Entier
— Name : Nom du passager. Chaîne de caractère
— Transported : Indique si le passager a été transporté avec succès. Booléen. Colonne cible.

2.2 Statistiques descriptives

La première étape de l'analyse exploratoire des données a été d'examiner les statistiques descriptives des caractéristiques numériques du jeu de données. Le tableau suivant récapitule des statistiques, qui fournissent un aperçu initial des tendances centrales, de la dispersion et de la forme de la distribution pour les variables Age, RoomService, FoodCourt, ShoppingMall, Spa, et VRDeck :

	Age	RoomService	FoodCourt	ShoppingMall	Spa	VRDeck
count	8514.000000	8512.000000	8510.000000	8485.000000	8510.000000	8505.000000
mean	28.827930	224.687617	458.077203	173.729169	311.138778	304.854791
std	14.489021	666.717663	1611.489240	604.696458	1136.705535	1145.717189
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	19.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	27.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	38.000000	47.000000	76.000000	27.000000	59.000000	46.000000
max	79.000000	14327.000000	29813.000000	23492.000000	22408.000000	24133.000000

FIGURE 2.1 – Résumé statistique

Nous pouvons observer que :

L'âge des passagers varie entre 0 et 79 ans, avec une moyenne d'environ 29 ans, ce qui suggère une population de passagers relativement jeune.

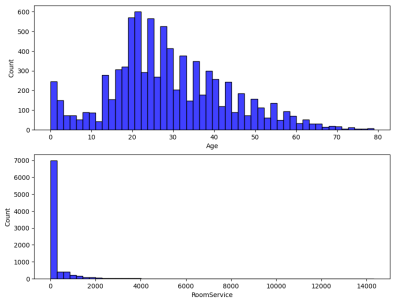
Pour les dépenses à bord, les moyennes des variables telles que RoomService, FoodCourt, ShoppingMall, Spa, et VRDeck indiquent que la majorité semble ne pas utiliser ces services, comme le montrent les médianes à zéro alors que certains passagers dépensent beaucoup.

L'écart-type élevé dans les dépenses des services à bord indique une grande variabilité, avec quelques passagers dépensant des montants assez élevés, affirmé par les valeurs maximales.

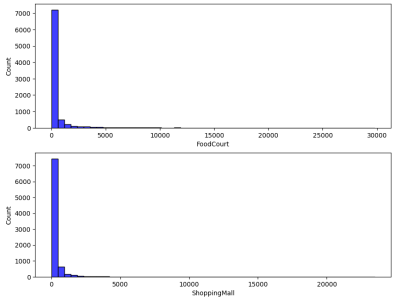
2.3 Distribution

2.3.1 Variables Quantitatives

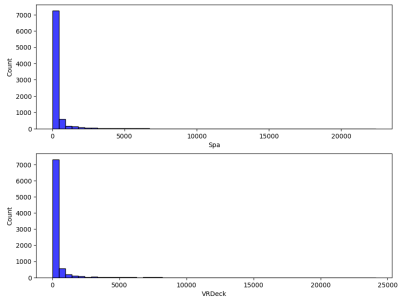
Pour comprendre les tendances et les comportements au sein du jeu de données, j'ai ploté des histogrammes qui illustrent les distributions pour les variables Age, RoomService, FoodCourt, ShoppingMall, Spa, et VRDeck. La distribution des âges présente une forme qui ressemble à une distribution normale avec un léger biais vers les jeunes adultes, avec un pic autour de la vingtaine. La présence de passagers de tous les âges, avec un âge maximum de 79 ans, suggère une diversité de la population à bord de la Spaceship Titanic.



(a) Distribution 'Age' et 'RoomService'



(b) Distribution 'FoodCourt' et 'ShoppingMall'



(c) Distribution 'Spa' et 'VRDeck'

FIGURE 2.2 – Distribution des variables quantitatives

Les variables RoomService, FoodCourt, ShoppingMall, Spa, et VRDeck montrent des distributions très basses pour dès que le montant dépasse 1000$, indiquant que la plupart des passagers dépensent peu ou pas dans ces services, tandis qu'une minorité dépense des sommes considérablement plus élevées. Cela est mis en évidence par les pics proches de zéro sur chaque histogramme et les valeurs maximales élevées, indiquant l'existence de dépenses exceptionnelles pour certains passagers.

2.3.2 Variables Qualitative : Transported

La variable qualitative Transported, la variable cible, présente une répartition presque égale entre les passagers transportés (True) et non transportés (False), comme l'indique le diagramme en barres. Cette équilibre dans la distribution de la variable cible indique qu'il n'y a pas de déséquilibre de classe significatif qui pourrait biaiser le modèle prédictif.

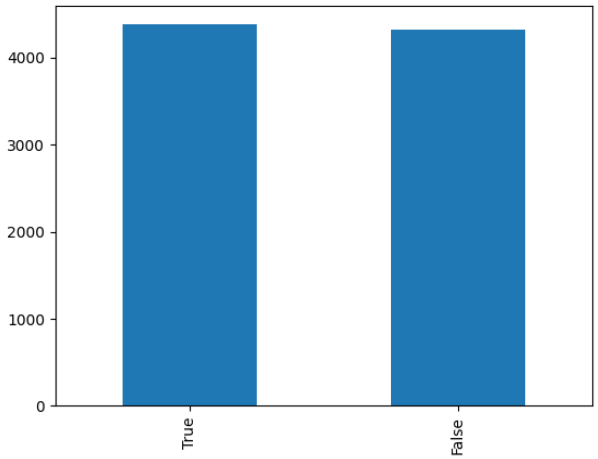


FIGURE 2.3 – Distribution 'Transported'

Dans la prochaine section, nous aborderons l'analyse des valeurs manquantes pour continuer à peaufiner notre compréhension des données.

Mais avant, dans le cadre des premières étapes de prétraitement, nous avons éliminé les colonnes PassengerId et Name du jeu de données. Ces attributs, ne contribuent pas de manière significative à la capacité prédictive du modèle en raison de leur nature non informative pour le processus de décision.

2.4 Valeurs manquantes

Le résumé suivant montre une quantification des valeurs manquantes :

— CryoSleep : 217 valeurs manquantes
— ShoppingMall : 208 valeurs manquantes
— VIP : 203 valeurs manquantes
— HomePlanet : 201 valeurs manquantes
— Cabin : 199 valeurs manquantes
— VRDeck : 188 valeurs manquantes
— FoodCourt : 183 valeurs manquantes
— Spa : 183 valeurs manquantes
— Destination : 182 valeurs manquantes
— RoomService : 181 valeurs manquantes
— Age : 179 valeurs manquantes
— Transported : 0 valeurs manquantes

La variable cible Transported ne présente aucune valeur manquante.

On observe une répartition relativement uniforme des valeurs manquantes à travers les variables, j'ai donc pu envisager une approche d'imputation. En effet pour les variables quantitatives, j'ai envisagé une imputation par la médiane pour préserver la distribution générale des données alors que l'auteur du Notebook a suggéré de les remplacer par une valeur nulle. Pour les variables qualitatives comme HomePlanet et CryoSleep, il a été suggéré dans le Notebook de laisser TF-DF s'occuper de l'imputation. La décision concernant Cabin, une variable qualitative avec une structure complexe, a nécessité une analyse supplémentaire pour déterminer la meilleure méthode d'imputation. En effet vu que cette colonne contient des chaîne de caractères sous le format Pont(Deck)/ Numéro de cabine/Coté, elle a été séparée en trois colonne prenant respectivement ces trois valeurs différentes. Après avoir extrait les informations nécessaires, la colonne Cabin originale du dataset a été supprimé car elle n'est plus nécessaire.

2.5 Corrélation

Après le traitements des valeurs manquantes j'ai voulu comprendre les interactions entre les différentes colonnes afin d'enrichir la compréhension des données. Alors que le Notebook ne le suggérait pas j'ai décidé de faire une analyse de corrélation qui est un outil essentiel pour comprendre les interdépendances potentielles entre les variables numériques de notre ensemble de données. J'ai utilisé une matrice de corrélation pour examiner l'association entre les diverses dépenses à bord et d'autres variables. Comme l'illustre la figure ci-dessous, la variable CryoSleep semble avoir une corrélation modérée et positive avec la variable cible Transported, ce qui suggère que les passagers en cryosommeil ont une probabilité plus élevée d'être transportés. D'autres variables de dépense présentent

des corrélations négatives avec CryoSleep, ce qui est cohérent avec l'idée que les passagers en sommeil cryogénique ne consomment pas de services à bord.

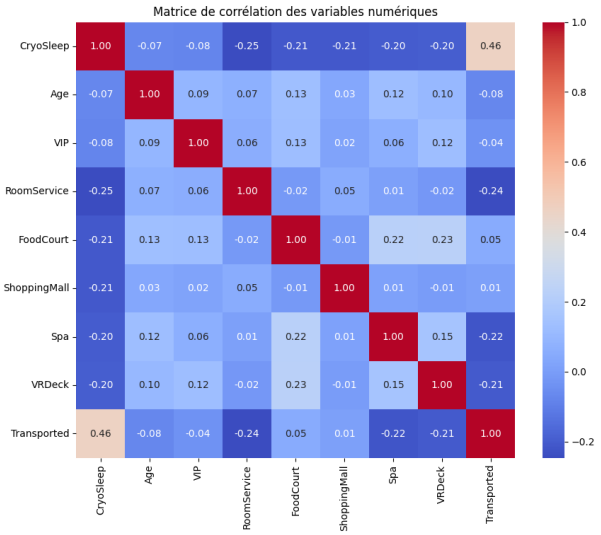


FIGURE 2.4 – Matrice de corrélation

L'analyse de la matrice de corrélation et des graphiques de distribution ont révélé des informations significatives sur les interactions entre les variables de notre jeu de données. La corrélation positive modérée entre CryoSleep et Transported souligne une caractéristique clé susceptible d'influencer la probabilité d'être transporté. Cette découverte visuelle fournit une base solide pour configurer le modèle prédictif.

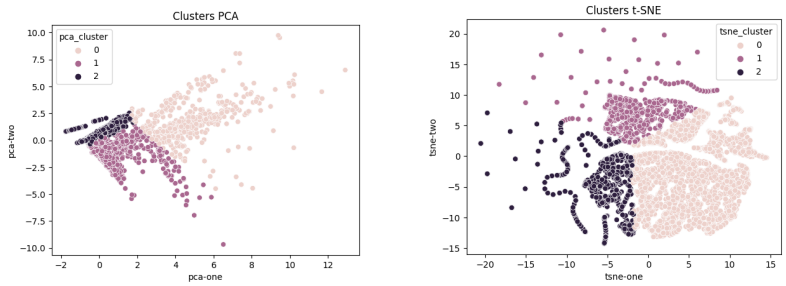
Afin d'enrichir encore plus la compréhension des données j'ai opté pour l'utilisation des techniques d'ACP et de t-SNE.

2.6 Analyse Multidimensionnelle

Les structures et groupements mis en évidence par des visualisations d'ACP et de t-SNE pourraient être exploités pour affiner les caractéristiques utilisées dans les modèles prédictifs ou pour identifier des sous-populations au sein des passagers. De plus ces analyses de visualisation multidimensionnelle sont particulièrement utiles pour identifier des variables clés qui influencent le comportement des passagers et la détection de schémas de données qui pourraient être pertinents pour la prédiction de la probabilité d'être transporté. Cependant, il convient de rester prudent quant aux limites de ces techniques, telles que la perte potentielle d'informations pertinentes lors de la réduction de dimensionnalité avec l'ACP, et la sensibilité du t-SNE aux hyperparamètres comme la perplexité.

2.6.1 Observations

Après exécution à l'aide de Scikit-learn, l'ACP et le t-SNE ont révélé des structures et des clusters distincts au sein du jeu de données. L'ACP a mis en évidence des profils de passagers variés, certains dépensant davantage dans des services tels que le FoodCourt et le VRDeck, suggérant un niveau de vie plus élevé ou des préférences spécifiques. D'autre part, le t-SNE a illustré des relations plus complexes et non linéaires entre les caractéristiques, avec des clusters suggérant des comportements ou des préférences sous-jacents moins évidents.



(a) Clustering sur l'ACP (b) Clustering sur le t-SNE

FIGURE 2.5 – Analyse Multidimentionnelle

Les passagers du Cluster 0, par exemple, montrent une tendance à être moins fréquemment 'transportés', ce qui auraient pu dû à des caractéristiques comportementales distinctes telles que l'évitement du CryoSleep ou un statut VIP.

2.7 Tests statistiques

Afin de vérifier la tendence observée dans la section précédente, j'ai effectué un test statistique dont l'hypothèse est de vérifier s'il existe une corrélation entre les dépenses en services, sachant que le statut VIP révèle des dépenses en services, et la probabilités d'être transporté. Pour cela la corrélation de Pearson entre les dépenses en RoomService (et similairement pour VRDeck) et la probabilité d'être transporté est de -0.241, avec une valeur p extrêmement basse (3.21e-115). La corrélation de Spearman est encore plus forte, à -0.347, avec une valeur p de 1.78e-244. Cela suggère que les passagers qui dépensent plus dans ces services sont moins susceptibles d'être transportés. La corrélation de Spearman, plus marquée que celle de Pearson, peut indiquer que cette relation est non linéaire.

Ces tests statistiques confirment que les habitudes de consommation, sont étroitement liés à la probabilité d'être transporté. Ces résultats sont essentiels pour le développement de notre modèle prédictif, car ils révèlent des variables potentiellement influentes qui

méritent une attention particulière. Toutefois, il convient de noter que ces relations ne prouvent pas la causalité et doivent être interprétées avec prudence.

Après cette exploration des données, qui a apporté une bonne compréhension des données, je vais aborder le chapitre suivant traçant la mise en place du modèle de prédiction. Cette étape va permettre d'appliquer les connaissances acquises précedemment afin de construire un modèle qui pourra prédire quel passager a été transporté ou non.

Chapitre 3

Modèle

Sommaire

3.1 Configuration et sélection du modèle

Avant de procéder à la configuration du modèle, les colonnes CryoSleep, VIP et Transported ont été convertis de booléens à entiers, comprenant les chiffres 1 ou 0. Cette conversion est nécessaire car TF-DF ne prend pas en charge les colonnes booléennes. Ensuite le dataset a été divisé avec une proportion de 80 /20 en deux datasets train_ds et valid_ds. Dans le Notebook, le modèle choisi de TF-DF est celui de RandomForestModel qui est l'algorithme d'entraînement de forêts de décision le plus connu. L'algorithme est censé être robuste face au suraprentissage et est aussi facile d'utilisation.

Afin d'entraîner le modèle j'ai opter pour la validation croisée à 5 plis afin d'évaluer la stabilité et la performance du modèle. Le jeu de données d'entraînement est divisé en plusieurs parties, dont un regroupement est un ensemble de tests et l'autre d'entraînements. Le modèle s'entraîne et est évalué à plusieurs reprises, chaque fois avec un nouvel ensemble. J'ai utilisé cette méthode afin de réduire la variance des estimations de performance et afin de juger la performance sur des données jamais vues. Son utilisation est robuste afin d'éviter des biais causés par un seul jeu de données de test.

Pour entraîner le modèle j'ai aussi appliqué une recherche des hyperparamètres les plus optimaux en procédant avec la méthode de Grid Search en variant le nombre d'arbres (50, 100, 200) et leurs profondeurs maximales(3, 5,10). J'ai choisi ces valeurs de façon arbitraire afin de commencer, sans prendre en compte le template d'hyperparamètres proposé par le Notebook « benchmark_rank1 » qui possèdent les hyperparamètres les plus optimaux dont le nombre d'arbres est égale à 300 et la valeur maximale de la profondeur vaut 16.

3.2 Résultats et discussions

Après exécution de l'entrainement du modèle j'ai obtenu les résultats suivant. Le modele le plus optimale selon ma recherche d'hyperparamètres est celui dont le nombre d'arbre est égale à 50 et dont valeur maximale de la profondeur vaut 10, ce qui est très différent du template « benchmark_rank1 » dont le nombre d'arbres est 6 fois plus grand et dont la profondeur maximale vaut 16. Après vérification des métriques, le modèle obtient des scores exceptionnelles, mais que j'ai fortement remis en question.
En effet :

- l'Accuracy vaut 0,9954,
- la Precision vaut 0,996,
- le Rappel vaut 0,9942,
- le F1 Score vaut 0,9953,
- et le ROC AUC Score qui vaut 0,9999.

De ce fait le modèle serait pratiquement parfait afin de distinguer quel passager a été transporté ou pas. De plus la génération de la matrice de confusion montre également des résultats surprenants, avec seulement 3 faux positifs et 5 faux négatifs.

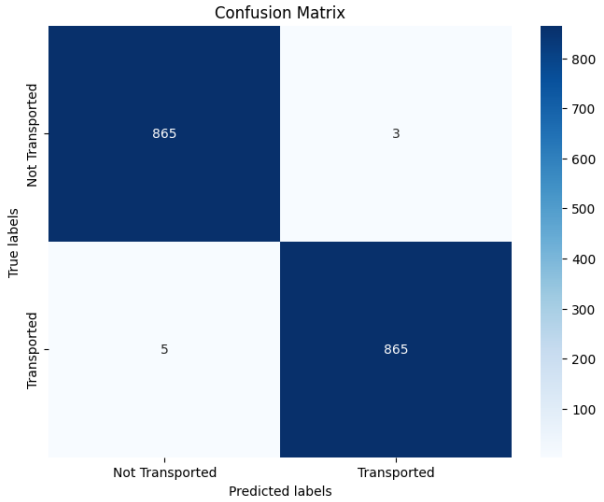


FIGURE 3.1 – Matrice de confusion'

Malgré ces résultats qu'on pourrait souligné de parfait j'ai suspecté un surajustement du modèle ou un sousapprentissage du modèle sur l'ensemble des données. En effet après avoir regardé les métriques j'ai évalué le modèle sur l'ensemble de donnée hors sac, Out Of Bag (OOB) qui permet de valider le modèle. Le modèle RFM choisit un échantillon de l'ensemble d'entraînement et le reste de l'echantillon est utilisé pour affiner le modèle. Le OOB est l'ensemble de donnée qui n'est pas utilisé à partir duquel le score est donc évalué.

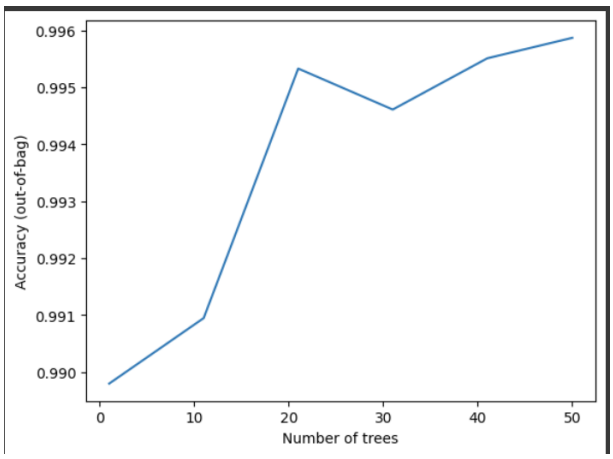


FIGURE 3.2 – Graphique OOB

Les scores obtenus pour l'ensemble OOB sont aussi surprenant que pour le jeu d'entraînement. J'ai obtenu un score de 0,9959. On observe sur le graphique une augmentation

significative au millième entre 10 et 20 arbre puis une baisse et enfin une légère croissance à 30 arbre jusqu'à 50 ou le modèle atteint le score obtenu. Le graphique montre aussi que le modèle n'est pas vraiment fiable car on y voit une fluctuation assez prononcée et la courbe ne se stabilise pas à la fin donc on a une certaine incertitude par rapport à comment elle va évoluer par la suite.

Et après observation sur l'ensemble de validation valid_ds j'ai obtenu une Accuracy de 0,9954. On pourrait croire qu'avec ces différents résultats de l'Accuracy que le modèle est idéal. Or pour affirmer cette excellence apparente, il a fallu exécuter le modèle sur l'ensemble de données test qu'il n'a jamais vu. D'abord il a fallu que je suivent toutes les étapes de prétraitement de données afin que l'ensemble convienne au modèle. Après exécution de ce dernier sur l'ensemble test, j'ai généré un fichier à soumettre sur Kaggle afin d'observer le score final obtenu. Mes suspicions se sont avérées justifiées car le modèle a obtenu un score de 0,39911, ce qui est nettement inférieur à l'Accuracy observées lors de la validation.

En réponse à cela, j'ai réadapté le modèle afin qu'il prenne le template d'hyperparamètres recommandé, « benchmark_rank1 » et après soumission j'ai obtenu une légère augmentation du score, qui est monté à 0.40011. Cette amélioration indique une meilleure, même si elle est petite, adaptation des hyperparamètres du template « benchmark_rank1 ».

Afin de comprendre les caractéristiques du modèle j'ai cherché à comprendre les variables déterminant celui-ci. En utilisant l'inspecteur du modèle, qui permet d'avoir un accès à la structure interne du modèle, j'ai pu observé les variables suivantes :

Variable	Importance	Catégorie et Référence
tsne-two	14.0	(1 ; #19)
pca-two	11.0	(1 ; #16)
pca_cluster	7.0	(1 ; #17)
CryoSleep	5.0	(1 ; #3)
pca-one	5.0	(1 ; #15)
tsne_cluster	5.0	(1 ; #20)
tsne-one	2.0	(1 ; #18)
RoomService	1.0	(1 ; #8)

TABLE 3.1 – Importance des variables du modèle

On peut voir que les variables importantes du modèle sont celles obtenues par la réduction de dimension des données. Mon but serait donc de déterminer lesquels de ces prédicteurs ont un effet disproportionné sur les prédictions du modèles. Avant d'effectuer cette recherche, j'ai cherché à savoir si cette réduction de dimension ne falsifiait justement pas le modèle en exécutant tout le code sans aucune réduction et en gardant le template « benchmark_rank1 ». Et après entraînement, test et soumission sur Kaggle, j'ai obtenue un score de 0,79354 pour une Accuracy de 0.7894 et 702 vrais négatifs, 693 vrais positifs, 189 faux négatifs et 183 faux positifs. Ce sont des résultats certes moins excellents pour

le modèle, or qui me donnent un score bien plus élevé pour le classement Kaggle.

Pour conclure, la réévaluation du modèle et une nouvelle exploration des données pour son amélioration sont inévitables. Cette exploration inclut une analyse des erreurs notamment celle des faux positifs et des faux négatifs. Elle inclurait aussi d'utiliser des techniques de régularisation comme le bagging ou boosting afin de contourner le surapprentissage, et aussi de chercher lesquelles des techniques de réduction de dimension sont incompatible avec le modèle et ses hyperparamètres.

Conclusion

En conclusion, ce mémoire a retracé la réalisation d'un projet traitant les étapes afin de développer un modèle prédictif dans le cadre d'une compétition Kaggle. Ce projet a permis d'appliquer de la théorie apprise lors des cours de la formation de Master MIASHS en effectuant une mise en pratique de cette théorie.

Il faut souligner qu'avant d'effectuer ce projet la première difficulté rencontrée a été de trouver une alternance. Un autre problème auquel j'ai dû faire face a été l'installation des bibliothèques spécifiques que j'ai surmonté en m'informant sur GitHub, les sites de ces bibliothèques, ou encore sur Internet. L'utilisation de Google Colab afin de régler les problèmes d'incompatibilités m'a montré un avantage d'utiliser des ressources cloud.

Le procédé pour mener ce projet à bien a été aligné aux objectifs du Master MIASHS. En effet j'ai pu exploiter les compétences en statistiques et en sciences des données. De plus, le fait d'avoir intégré l'apprentissage automatique et la visualisation des données a donné matière sur le potentiel de ces outils pour la compréhension d'un grand ensemble de données, qui est une aptitude recherchée par le Master.

N'étant pas lié à une situation d'alternance en entreprise, je n'ai pas eu à faire face à une responsabilité sociétale des entreprises (RSE) ou au développement durable et responsabilité sociale (DDRS). Sachant que les données du projet sont fictives au vu de son contexte, il n'y a aucune situation d'exploitation de données sensibles ou privés.

Pour perspective, ce projet pourra m'apporter en termes de mise en pratique des méthodes d'analyse de données et il pourra aussi devenir une sorte de préparation dans le monde professionnel notamment dans le domaine de la data science, avec une bonne compréhension de l'importance d'une validation externe de modèles prédictifs. En somme, le mémoire a démontré l'utilisation en pratique quelques compétence acquises au cours du Master MIASHS. De plus lors de la réalisation de ce projet, j'ai pu expérimenté l'importance d'être flexible, d'être créatif et d'avoir une réflexion critique dans le domaine de la science des données. L'objectif final étant de comprendre comment on procède à la construction d'un modèle performant, stable et généralisable, en partant de l'exploration d'un jeu de données et en passant par leur prétraitement, dont la finalité dans le contexte de ce projet est de pouvoir sauver des passagers qui seraient perdus dans une dimension parallèle.

Bibliographie

[1] Ryan Holbrook Addison Howard, Ashley Chow. Spaceship titanic, 2022.

[2] eliot robot Gusthema. spaceship-titanic-with-tfdf.ipynb, 2022.

[3] Wes McKinney. Pandas.

[4] Travis E. Oliphant. Numpy.

[5] Michael Waskom. Seaborn.

[6] Michael Droettboom John Hunter. Matplotlib.

[7] Forum : Kaggle competition spaceship titanic. Forum de discussion.

[8] Serveur discord de kaggle. Serveur de discussion.

[9] Stackoverflow. Pour l'explication d'erreurs, debugging, et propositions de solutions de code.

[10] Chatgpt. Pour l'explication d'erreurs, debugging, propositions de solutions de code, et éléments d'interprétations.

[11] Gwenaël Richomme Catherine Trottier Sandra Bringay, Sophie Lèbre. Introduction à la science des données. Application d'éléments du cours.

[12] Pierre Lafaye De Micheaux. Analyse de données multidimensionnelles. Application d'éléments du cours.

[13] Sophie Lèbre Marine Demangeot. Classification supervisée et non supervisée. Application d'éléments du cours.