[?]

UFR 6
Université Paul Valéry, Montpellier III

Rapport de projet

Kaggle Competition : Spaceship Titanic

Jonathan Duckes

Mai 2024

Remerciements

Je souhaite exprimer ma gratitude envers les responsables de la formation, Madame Sophie Lèbre, Monsieur Maximilien Servajean et Monsieur Arnaud Sallabery, pour leur accompagnement, leur disponibilité et surtout pour m'avoir offert la chance de travailler sur un projet en raison de ma situation exceptionnelle.
J'aimerais également exprimer ma reconnaissance envers mes camarades et collègues de classe, qui m'ont donné des conseils et des jugements pertinents pour progresser dans ce projet.
Je tiens également à exprimer ma gratitude envers l'équipe pédagogique et les enseignants de la formation pour leur enseignement et leur transmission des outils et des connaissances qui ont été d'une grande aide pour mener à bien ce projet.

Résumé

Les étudiants peuvent bénéficier d'une expérience pratique en alternance en Master MIASHS, qui leur permet d'appliquer les mathématiques et l'informatique aux sciences humaines et sociales. Il s'agit de renforcer les compétences professionnelles des étudiants et de leur donner l'opportunité de mettre en pratique leurs connaissances théoriques. En général, les alternances portent sur l'exploration et l'analyse de données, la modélisation mathématique ou l'utilisation d'outils informatiques pour résoudre des problèmes réels. Les étudiants peuvent ainsi s'initier au monde du travail et prendre part à des projets de recherche en cours. Ils bénéficient de l'accompagnement d'un tuteur tout au long de l'alternance.

Malheureusement, je n'ai pas pu trouver d'alternance pour ma première année de Master MIASHS, mais grâce à l'aide précieuse de mes professeurs, j'ai réussi à trouver une alternative constructive. Ils m'ont offert la possibilité de collaborer sur un projet spécifique. Effectivement, ils m'ont demandé de découvrir un projet Kaggle [1], une plateforme spécialement conçue pour les chercheurs en données et l'apprentissage automatique. La plateforme offre la possibilité de prendre part à des concours de Machine Learning et de mener à bien des projets. Ainsi, Kaggle est extrêmement bénéfique pour mettre en pratique les connaissances en apprentissage automatique.

Dans ce mémoire, je vais présenter les visualisations et les statistiques descriptives que j'ai développées pour comprendre et analyser le jeu de données afin de satisfaire aux exigences de mes professeurs. Il sera également inclus dans ce mémoire des descriptions des solutions statistiques et de machine learning mises en place pour obtenir un modèle prédictif, ainsi que des descriptions, des métriques, des tâches, du contexte, etc.

Table des matières

Table des figures

Liste des tableaux

Introduction

L'UFR 6 de l'Université Paul-Valéry Montpellier 3 propose un Master MIASHS qui enseigne aux étudiants le métier de data scientist, qui se spécialise dans les données "massives et complexes" et les méthodes d'apprentissage statistique [2]. En raison de sa nature pluridisciplinaire, la formation allie mathématiques, informatique et science humaine pour donner aux étudiants une perspective globale sur le traitement de données.La science des données joue un rôle crucial dans la compréhension et la modélisation des problématiques professionnelles. Le rôle du data scientist est crucial dans le processus de traitement des données, de collecte et d'interprétation de ces dernières.

L'objectif de ce mémoire est d'explorer les activités que j'ai effectuées pendant ma première année en Master MIASHS, de mettre en évidence les diverses compétences que j'ai développées et perfectionnées grâce à un projet instructif et captivant, et enfin de proposer des perspectives futures pour mener à bien le projet présenté. Plus spécifiquement, ce mémoire portera sur une compétition Kaggle intitulée « Spaceship Titanic » [3], dont le but était de prédire quels passagers ont été transportés dans une dimension alternative lors d'une collision avec une anomalie spatio-temporelle.

Chapitre 1

Contexte et Objectifs du projet

Sommaire

1.1 Contexte

Nous nous trouvons en 2912 où le Spaceship Titanic, un vaisseau spatial dérivé du bateau Titanic, transportant environ 13000 passagers, parcourt notre système solaire afin de les amener sur trois nouvelles planètes habitables. Malheureusement, tout comme son ancêtre d'il y a 1000 ans, il rencontre un destin similaire. En effet, le Spaceship Titanic entre en collision avec une anomalie spatio-temporelle, ce qui provoque la téléportation de près de la moitié des passagers dans une dimension alternative. L'objectif de notre mission est de prédire les passagers qui ont été transporté afin d'aider l'équipe de secours à les retrouver. Cette tâche de classification s'appuie sur divers attributs des passagers et de leur voyage.

1.2 Présentation des données

Les données fournies pour cette compétition sont composées des fichiers suivants :

- **train.csv** : Enregistrements personnels pour environ deux tiers des passagers (~8700), utilisés comme ensemble d'entraînement.
  - **PassengerId** : Un identifiant unique pour chaque passager.
  - **HomePlanet** : La planète d'origine du passager ou la planète de départ.
  - **CryoSleep** : Indique si le passager a opté pour une animation suspendue pendant le voyage.
  - **Cabin** : Le numéro de cabine du passager. Prend la forme deck/num/side (pont/num/côté).
  - **Destination** : La planète de destination du passager.
  - **Age** : L'âge du passager.
  - **VIP** : Indique si le passager bénéficie du statut VIP.
  - **RoomService, FoodCourt, ShoppingMall, Spa, VRDeck** : Montants facturés aux différentes installations de luxe du Spaceship Titanic.
  - **Name** : Les prénoms et noms des passagers.
  - **Transported** : Indique si le passager a été transporté dans une dimension alternative. C'est la cible à prédire.
- **test.csv** : Enregistrements personnels pour le tiers restant des passagers (~4300), utilisés comme ensemble de test.
- **sample_submission.csv** : Un fichier de soumission exemple au format correct.

Chapitre 2

Méthodologie de Gestion de projet

Sommaire

2.1 Méthodologie de Gestion de Projet

Pour mener à bien ce projet, j'ai adopté une méthodologie structurée en suivant les différentes étapes et en utilisant les supports de cours dispensés durant mon Master MIASHS. Voici un aperçu de cette méthodologie :

2.1.1 Approche de gestion de projet

Planification des étapes du projet

— La première étape a consisté à revoir et comprendre les cours pertinents pour le projet, notamment les cours de « Régression logistique et poissonienne » [4], « Théorie du Machine Learning » et « Classification supervisée et non supervisée » [5]. Cette phase de révision a permis de solidifier les bases théoriques nécessaires à l'implémentation pratique.
— Ensuite, j'ai planifié la réalisation du projet en divisant les tâches en différentes phases : exploration des données, préparation des données, implémentation des modèles, et évaluation des modèles.

Répartition des tâches

— **Exploration des données** : Cette phase comprenait la lecture du dataset, la vérification des valeurs manquantes et une première visualisation des distributions des variables.
— **Préparation des données** : Nettoyage des données, gestion des valeurs manquantes, transformation des variables catégorielles en variables numériques.
— **Implémentation des modèles** : Création et entraînement de différents modèles de classification, tels que la régression logistique, le Random Forest et le SVM.
— **Évaluation et sélection des modèles** : Comparaison des performances des modèles à l'aide de métriques telles que l'accuracy, la courbe ROC et l'AUC.
— **Génération des prédictions finales et soumission** : Application du modèle sélectionné sur les données de test et création du fichier de soumission pour Kaggle.

2.1.2 Outils et techniques de gestion de projet utilisés

— **Jupyter Notebook** [6] : Pour l'implémentation du code Python, l'exploration des données et la visualisation.
— **Microsoft To Do** [7] : Pour avoir un suivi des tâches.
— **Documentation des cours** : Utilisation des supports de cours pour guider l'implémentation et la compréhension des méthodes.

Chapitre 3

Analyse des Données

Sommaire

3.1 Exploration des données

Pour ce projet, j'ai effectué une exploration des données approfondie afin de mieux comprendre leurs caractéristiques.

3.1.1 Statistiques descriptives

La première étape de cette exploration a été d'observer les statistiques descriptives des variables numériques. Les tableaux ci-dessous montrent les premières lignes du jeu de données, les statistiques descriptives et un aperçu des informations sur les colonnes.

	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 3.1 – Résumé statistique

Ces statistiques nous donnent une vue d'ensemble sur la distribution, les tendances et la dispersion des variables Age, RoomService, FoodCourt, ShoppingMall, Spa, et VRDeck :

Nous observons que :

— L'âge des passagers se situe entre 0 et 79 ans, avec une moyenne aux alentours de 29. Nous pouvons en conclure que les passagers à bord du vaisseau sont jeunes.

— Les médianes à 0 des variables de dépense suggèrent qu'une majorité des passagers n'utilise pas les services, or leur moyenne indique que d'autres dépensent beaucoup. L'écart-type de ces dépenses montre une variabilité élevée sachant qu'il y a de hautes valeurs maximales.

3.1.2 Visualisations des distributions des variables

Afin de mieux comprendre la distribution des variables, j'ai utilisé des diagrammes en barres qui illustrent la distribution des variables Age, RoomService, FoodCourt, ShoppingMall, Spa, VRDeck et Transported.

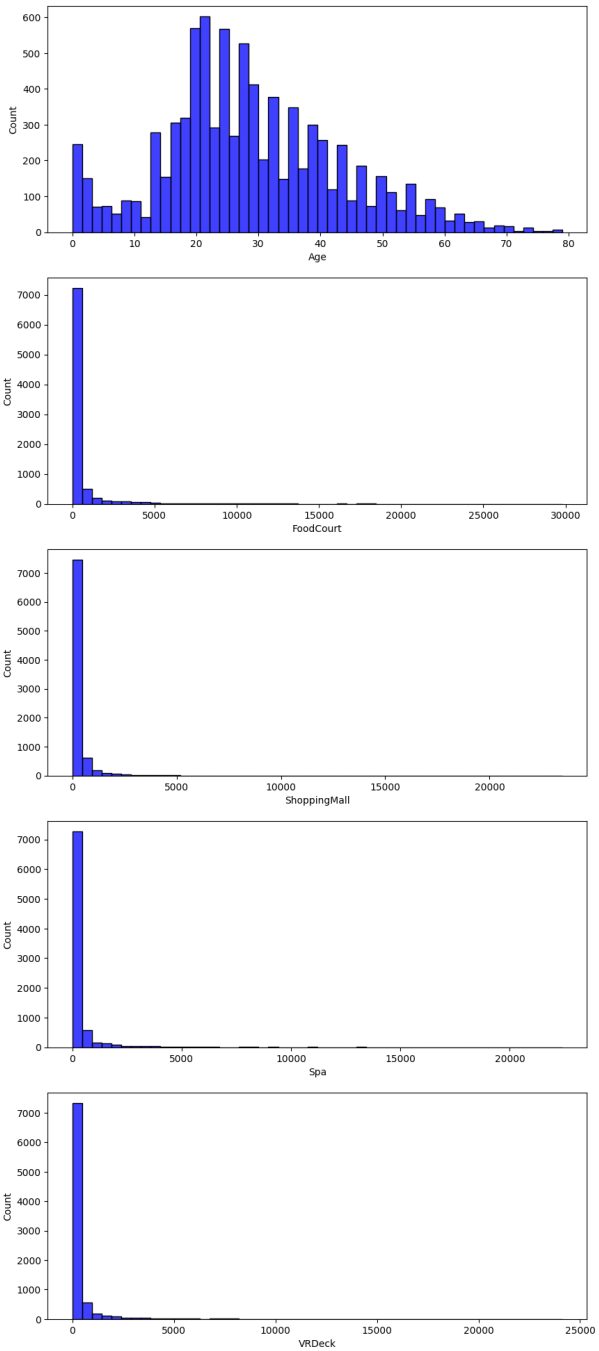


FIGURE 3.2 – Distribution des variables *Age*, *RoomService*, *FoodCourt*, *ShoppingMall*, *Spa* et *VRDeck*.

L'histogramme de Age montre une certaine diversité des passagers. En effet, nous remarquons ici que l'âge maximum est 79, comme présenté dans les statistiques descriptives, et nous observons aussi un pic de distribution autour de la vingtaine.

Les histogrammes des variables des dépenses affichent des tendances similaires, nous voyons clairement des pics proches de zéro ce qui affirme le fait que la plupart des passagers dépensent peu, de plus nous observons des distributions très basses pour des valeurs dépassant 1000$.

La distribution de la variable Transported montre un équilibre avec une répartition presque égale entre les passagers qui ont été transportés (True) ou non (False), ce qui est illustré dans le diagramme en barre ci-dessous.

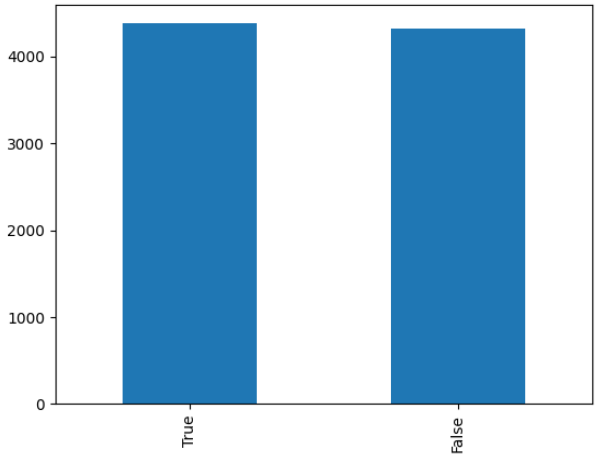


FIGURE 3.3 – Distribution de la variable *Transported*.

L'analyse des statistiques descriptives et des visualisations des variables a permis une compréhension de base des données. La distribution des âges, par exemple, montre que la majorité des passagers sont jeunes, et celle des dépenses suggère que l'utilisation des services du vaisseau est quasiment nulle pour l'ensemble des passagers, à quelques exceptions.

3.2 Traitement des valeurs manquantes

Avant de commencer, j'ai suivi l'idée du Notebook *Spaceship Titanic with TFDF* [8] de 'Gusthema' et 'eliot robot' [], qui est de préparer la colonne Cabin. Celle-ci a été séparée en trois nouvelles colonnes « Deck » (pont), « Cabin_num » (numéro de cabine) et « Side » (côté). Ensuite, la colonne « Cabin » a été supprimée.

Pour l'étape de la gestion des valeurs manquantes, j'ai d'abord identifié les colonnes comportant des valeurs manquantes dans le jeu de données.

Les valeurs manquantes des variables qualitatives, « HomePlanet », « CryoSleep », « Destination », « VIP », « Deck », « Cabin_num » et « Side » ont été remplacées par la valeur la plus fréquente.

Les valeurs manquantes pour la variable « Age » ont été remplacées par la médiane des âges.

Les colonnes des dépenses « RoomService », « FoodCourt », « ShoppingMall », « Spa », et « VRDeck » ont été remplies par des zéros, car l'absence de valeur peut être interprétée comme une absence de dépense.

Les variables catégorielles ont ensuite été encodées en utilisant des variables numériques afin de pouvoir être utilisées dans les modèles.

Après le traitement des valeurs manquantes, seules les colonnes « Name » avaient encore des valeurs manquantes, qui ont été ignorées car elles n'étaient pas pertinentes pour les modèles prédictifs.

3.3 Visualisations et interprétations

Pour mieux comprendre les relations entre les variables et leurs impacts potentiels sur l'objectif de prédiction, plusieurs visualisations ont été réalisées.

3.3.1 Matrice de Corrélation des Variables

La matrice de corrélation des variables numériques est présentée dans la première visualisation. Les relations linéaires entre les différentes variables du jeu de données peuvent être identifiées grâce à cette matrice. Le coefficient de corrélation de Pearson [9] est utilisé pour évaluer la corrélation, avec une valeur allant de -1 à 1.

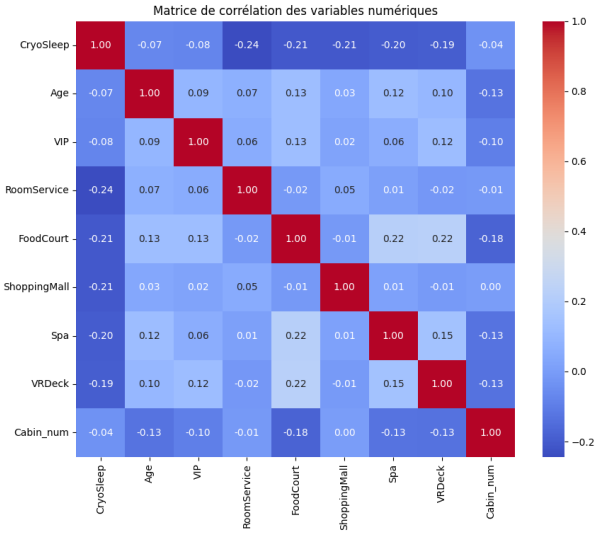


FIGURE 3.4 – Matrice de corrélation des variables numériques du dataset. Les couleurs représentent les coefficients de corrélation de Pearson, allant de -1 (corrélation négative maximale) à 1 (corrélation positive maximale).

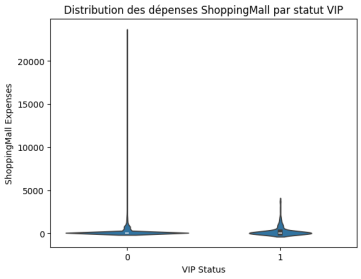
1. **CryoSleep** : Des liens négatifs sont observés avec les dépenses (*RoomService, FoodCourt, ShoppingMall, Spa, VRDeck*). Cela laisse entendre que les passagers en CryoSleep ont tendance à économiser davantage dans ces services, ce qui est compréhensible si l'on suppose qu'ils sont en hibernation et ne peuvent donc pas les utiliser.
2. **Les dépenses** : Il y a une corrélation positive entre les différentes variables de dépenses. Par exemple, il existe une corrélation modérée entre les dépenses au *FoodCourt* et au *Spa* (0.22). Cela pourrait suggérer que les voyageurs qui dépensent dans une catégorie ont également tendance à dépenser dans une autre catégorie.
3. **L'âge** présente de faibles corrélations avec d'autres variables, ce qui suggère que l'âge des passagers n'a pas un impact significatif sur leur comportement de dépenses ou leur décision d'être en CryoSleep.

3.3.2 Analyse des Dépenses en Fonction du Statut VIP

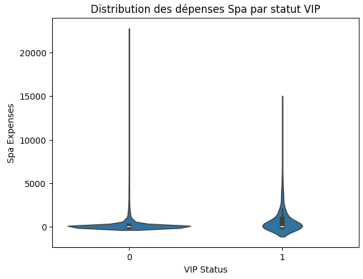
Pour analyser l'impact du statut VIP sur les dépenses, des boxplots ont été générés pour chaque type de dépense en fonction du statut VIP.

Interprétation :

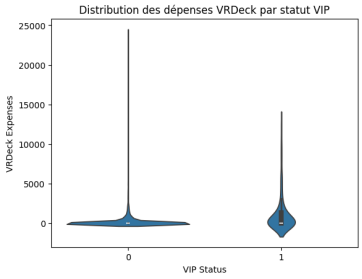
1. **ShoppingMall, Spa et VRDeck** : Ces services présentent des tendances similaires. En règle générale, les passagers VIP dépensent davantage, mais ces dépenses sont encore une fois ciblées sur des valeurs basses.



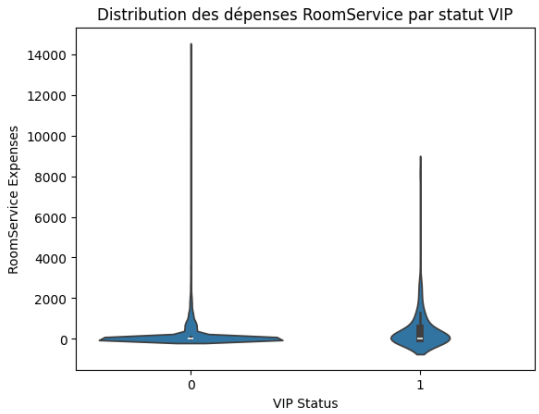
(a) Dépenses au ShoppingMall par statut VIP



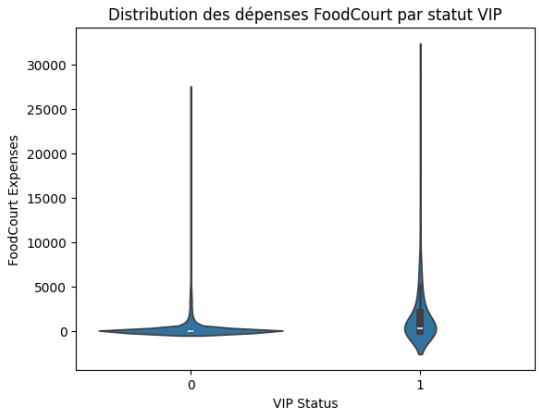
(b) Dépenses au Spa par statut VIP



(c) Dépenses au VRDeck par statut VIP



(d) Dépenses au RoomService par statut VIP



(e) Dépenses au FoodCourt par statut VIP

FIGURE 3.5 – Comparaison des dépenses des passagers VIP et non-VIP dans différents services du Spaceship Titanic

2. **RoomService** : Les passagers VIP consacrent davantage de fonds au service de chambre que les passagers non-VIP, même si la plupart des dépenses restent concentrées sur des valeurs faibles pour les deux groupes.
3. **FoodCourt** : Il est également observé que les passagers VIP ont tendance à dépenser davantage que les non-VIP, mais tout comme pour le *RoomService*, la plupart des passagers ne dépensent pas ou peu.

De plus, le nombre de passagers VIP et non-VIP est très disproportionné, avec seulement 2,3 % des passagers ayant un statut VIP. La faible proportion de VIP pourrait être à l'origine de la faible corrélation entre le statut VIP et les variables de dépenses observées précédemment.

3.3.3 Conclusion des Visualisations et Interprétations

Les comparaisons et les représentations visuelles des dépenses en fonction du statut VIP offrent des renseignements essentiels sur les comportements des passagers du Spaceship Titanic. Malgré la prédisposition des passagers VIP à dépenser davantage dans les divers services du vaisseau, la plupart des passagers ne dépensent pas ou très peu, ce qui laisse entendre des comportements économiques divers à bord du vaisseau. Il est essentiel de prendre en compte ces observations afin de développer et d'améliorer les modèles de prédiction utilisés pour évaluer les passagers transportés.

Chapitre 4

Modélisation

Sommaire

4.1 Modélisation et Évaluation des Modèles

4.1.1 Choix des Modèles de Machine Learning

Dans le cadre de ce projet, trois modèles principaux ont été sélectionnés en fonction de leur pertinence et des cours suivis pendant le semestre : Le premier a été la régression logistique, car celle-ci est utilisée pour prédire une variable binaire. Le modèle estime la probabilité qu'une observation appartienne à une certaine classe. Le deuxième modèle utilisé a été un Random Forest qui combine plusieurs arbres de décision pour améliorer la précision. Ce modèle est très robuste contre le surapprentissage et possède une capacité à gérer des données complexes et non-linéaires. Le dernier modèle que j'ai choisi est le SVM car celui-ci cherche à trouver l'hyperplan qui maximise la marge entre les différentes classes.

Avant de mettre en place ces modèles, j'ai choisi d'intégrer toutes les variables explicatives dans la matrice X, à l'exception du numéro d'identifiant des passagers et de leur âge, car ces informations ne sont pas essentielles pour la modélisation. Ensuite, l'ensemble des données d'entraînement a été divisé en train-test avec une proportion de 80/20. Par la suite, j'ai également standardisé les données pour garantir une convergence rapide du modèle.

4.2 Résultats et discussions

Pour le modèle de régression logistique, j'ai décidé d'appliquer les méthodes de mon cours, c'est-à-dire que j'ai configuré un modèle ne prenant aucune variable explicative afin de le comparer avec un modèle prenant toutes les variables explicatives. La comparaison des deux modèles s'est faite par le calcul de la log-vraisemblance des deux modèles afin de faire un test du rapport de vraisemblance pour vérifier si le modèle complet est plus significatif que le modèle nul. Mon test a bien montré que le modèle complet est plus significatif que le modèle nul. En créant la table de classification suivante, nous observons que le modèle complet a un taux de bonne classification de 78.38%, avec une sensibilité de 82.23% et une spécificité de 74.45%, ce qui signifie que le modèle est globalement précis pour prédire les passagers transportés et non transportés, avec une légère meilleure performance pour identifier correctement les passagers transportés.

	Prédit Non Transporté	**Prédit Transporté**
Réel Non Transporté	641	220
Réel Transporté	156	722

TABLE 4.1 – Table de classification

— **Taux de bonne classification** : 0.7838
— **Sensibilité** : 0.8223
— **Spécificité** : 0.7445

Ensuite, pour évaluer les performances du modèle, j'ai ploté l'évolution de la sensibilité et de la spécificité en fonction d'un seuil c afin de voir sa valeur optimale. On peut voir sur la courbe ci-dessous que la valeur optimale est 0,56.

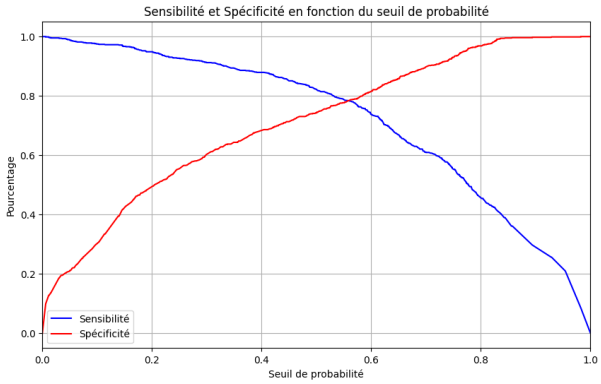


FIGURE 4.1 – Évolution de la sensibilité et de la spécificité en fonction du seuil de probabilité.

Suite à ces résultats, je me suis demandé si je pouvais sélectionner un modèle contenant moins de variables explicatives et si celui-ci pourrait être plus performant. J'ai fait cette comparaison à l'aide des critères BIC et AIC. À partir du modèle complet, j'ai comparé son AIC et BIC avec un modèle contenant une variable en moins et j'ai gardé le modèle ayant un AIC et BIC inférieur. Après avoir tourné mon algorithme, le modèle ayant un AIC et BIC favorable était celui contenant les variables « CryoSleep », « RoomService », « FoodCourt », « Spa », « VRDeck », « HomPlanet_Europa », « Deck_C », « Deck_G » et « Side_S ». Pour évaluer ce modèle, j'ai regardé la courbe ROC, qui illustre la performance du modèle en regardant son taux de positifs par rapport à son taux de négatifs. L'aire sous la courbe AUC, qui mesure la capacité du modèle à distinguer entre les classes, obtenue est de 0.86, ce qui veut dire que la discrimination est excellente, donc le modèle arrive assez bien à déterminer si un passager a été transporté ou non.

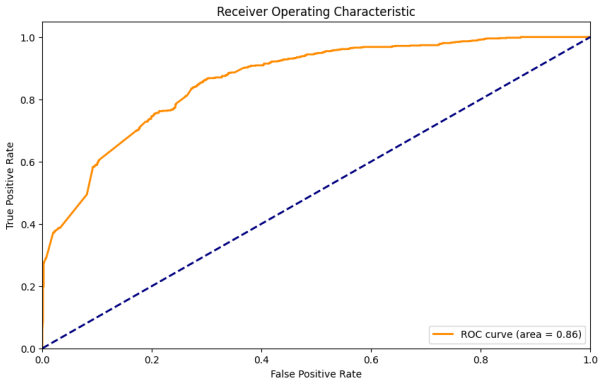


FIGURE 4.2 – Courbe ROC (Receiver Operating Characteristic) pour le modèle de régression logistique avec une aire sous la courbe (AUC) de 0.86. La courbe ROC illustre la capacité du modèle à distinguer entre les classes. Plus l'aire sous la courbe est proche de 1, meilleure est la performance du modèle.

Après avoir fait cette exploration avec le modèle de régression logistique, l'idée a été de sélectionner un modèle dont la performance est la meilleure possible sachant que la vie de certains passagers est en danger. Pour ce faire, j'ai donc choisi les modèles de Random Forest et le SVM pour les raisons citées plus haut et j'ai fait la comparaison des performances en observant la courbe ROC et l'aire en dessous de la courbe AUC.

D'abord, j'ai configuré le modèle de Random Forest dont j'ai évalué les performances pour différents nombres d'arbres, de 1 à 50, avec une fonction de coût 0/1 afin de déterminer le nombre d'estimateurs

offrant la meilleure précision, comme fait dans un TP [10] dans le cours de « Régularisation et optimisation ». J'ai ensuite observé son taux d'erreur dans l'ensemble train et ensuite dans l'ensemble test en fonction du nombre d'estimateurs. Nous pouvons observer ci-dessous que la courbe d'erreur d'entraînement (en bleu) diminue rapidement pour atteindre un niveau très bas, indiquant un ajustement très précis aux données d'entraînement. En revanche, la courbe d'erreur de test (en rouge) diminue initialement, mais se stabilise autour de 0.22, suggérant que l'ajout d'estimateurs supplémentaires n'améliore pas significativement la performance du modèle sur les données de test, ce qui peut indiquer un début de surapprentissage.

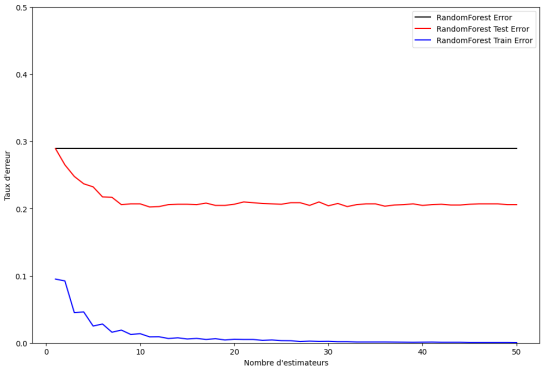


FIGURE 4.3 – Évolution du taux d'erreur en fonction du nombre d'estimateurs pour le modèle Random Forest.

Pour finir, j'ai tracé la courbe ROC de ce modèle dont l'AUC a pour valeur 0.875, ce qui est plus élevé que l'AUC du modèle de régression logistique.

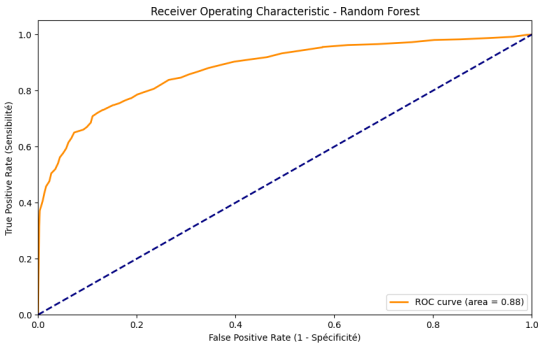


FIGURE 4.4 – Courbe ROC (Receiver Operating Characteristic) pour le modèle de régression logistique avec une aire sous la courbe (AUC) de 0.875.

Pour le dernier modèle, vu que les données ne sont pas linéairement séparables, j'ai configuré un SVM à noyau RBF dont j'ai défini le paramètre de régularisation C à 1 pour débuter. Pour évaluer ce modèle, j'ai regardé son taux de bonne classification qui vaut 0.788 et j'ai aussi tracé sa courbe ROC dont l'AUC a pour valeur 0.878, ce qui est meilleur que les deux précédents modèles.

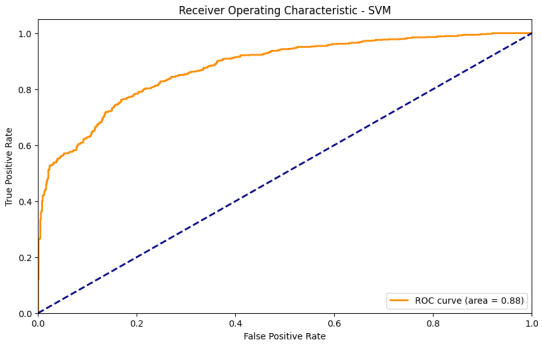


FIGURE 4.5 – Courbe ROC (Receiver Operating Characteristic) pour le modèle de régression logistique avec une aire sous la courbe (AUC) de 0.878.

Chapitre 5

Soumission

Sommaire

5.1 Soumission Kaggle

Après configuration et évaluation de ces modèles, j'ai sélectionné le modèle SVM car son taux de bonne classification et l'AUC de sa courbe ROC étaient les plus élevés des trois modèles. Suite à cela, il a fallu tester le SVM sur l'ensemble test fourni par Kaggle afin de le soumettre et d'observer le score obtenu.

D'abord, il a fallu faire les mêmes étapes de prétraitement que pour l'ensemble d'entraînement, donc gérer les valeurs manquantes de chaque colonne de la même façon expliquée dans la section « Traitement de données manquantes » et normaliser les variables explicatives. Ensuite, j'ai exécuté le modèle sur l'ensemble de données test qui a donc prédit si un passager a été transporté ou non. J'ai ensuite créé le fichier de soumission en reprenant le sample fourni par Kaggle. Après soumission de celui-ci sur Kaggle, j'ai obtenu un score de 0.79939, ce qui est plus élevé que les scores que j'avais obtenus lors du premier semestre avec un Notebook de base de Kaggle (« How to get started » [8]).

FIGURE 5.1 – Capture d'écran montrant le score de soumission sur Kaggle avec un modèle SVM, obtenant un score de 0.79939.

Chapitre 6

Organisation et Prise de Recul

Sommaire

6.1 Gestion de projet

Pour réaliser ce projet, j'ai adopté une approche agile pour organiser les diverses étapes de travail. J'ai employé des outils comme Microsoft To Do afin de planifier et de suivre les tâches à effectuer, ce qui m'a aidé à maintenir mon ordre et à respecter les échéances. Les cours de gestion de projet dispensés dans le cadre du Master MIASHS ont été l'inspiration de la méthodologie.

Toutefois, en prenant du recul, j'aurais pu accomplir davantage de tâches. Chaque fois que je travaillais, j'avais deux semaines pour progresser sur le projet. Il m'est arrivé de penser avoir assez de temps, ce qui m'a poussé à ne pas être assez exigeant dans mon travail. Cette expérience m'a enseigné l'importance de maintenir un rythme de travail constant, notamment en milieu professionnel, où il est essentiel de présenter des résultats concrets dans les délais impartis.

6.2 Retour d'expérience

Au cours du projet, divers obstacles ont été relevés, tels que la gestion des valeurs manquantes et la sélection des modèles. À titre d'exemple, j'ai divisé la colonne 'Cabin' en trois nouvelles colonnes (Deck, Cabin_num, Side) pour faciliter la gestion des informations manquantes et faciliter l'analyse. En outre, j'ai employé des méthodes de régularisation afin d'améliorer les performances et la disponibilité des modèles généralisés.

J'ai pu améliorer mes compétences en analyse de données, en programmation (Python) et en machine learning grâce à ce projet. Ma compréhension des techniques de prétraitement des données, de modélisation et d'évaluation de modèles prédictifs s'est considérablement améliorée. J'ai aussi développé des aptitudes en gestion de projet et en planification de tâches, ce qui m'a permis de mieux structurer mon travail et de respecter les échéances.

6.3 Apport du master MIASHS

Les cours dispensés lors du Master MIASHS ont joué un rôle crucial dans la concrétisation de ce projet. Ce projet a directement mis en pratique les concepts théoriques acquis, tels que la régression logistique, les forêts aléatoires, les SVM et les méthodes de régularisation. L'analyse et la modélisation des données ont également nécessité des compétences en programmation et en utilisation de bibliothèques Python (Pandas [11], Numpy [12], Scikit-learn [13], Matplotlib [14], Seaborn [15]).

Chapitre 7

Perspectives

Sommaire

7.1 Perspectives futures

Afin de continuer ce projet, différentes étapes sont envisagées :

— **Amélioration des modèles** : Recherche de nouvelles méthodes d'apprentissage automatique et optimisation des hyperparamètres afin d'améliorer la précision des prédictions.

— **Analyse détaillée des erreurs** : Analyser les fausses positives et les fausses négatives afin de saisir les contraintes des modèles actuels et repérer les améliorations envisageables.

— **Création de nouvelles caractéristiques** : Incorporation de variables explicatives supplémentaires ou modification des variables déjà présentes afin d'enrichir le modèle.

À l'avenir, je prévois d'améliorer mes compétences en machine learning avancé et en big data, ainsi que de participer à d'autres compétitions Kaggle afin de poursuivre mon apprentissage et mon utilisation de nouvelles techniques et aussi d'intégrer une entreprise pour suivre ma formation en alternance. J'envisage également de développer des compétences en gestion de projet et en collaboration afin d'améliorer l'efficacité et la collaboration dans des projets à venir.

Le métier de data scientist connaît une évolution rapide en raison de l'augmentation des quantités de données et de l'apparition de nouvelles technologies. Il est de plus en plus essentiel d'avoir la capacité d'analyser et d'interpréter des données complexes. Les data scientists doivent aussi posséder des compétences en communication et en gestion de projet afin de travailler de manière efficace avec des équipes interdisciplinaires. Il est probable que l'évolution future du métier impliquera une automatisation accrue des tâches de science des données et une intégration plus étroite avec les domaines de l'intelligence artificielle et de l'apprentissage automatique.

Chapitre 8

Conclusion

Sommaire

8.1 Conclusion

8.1.1 Résumé des réalisations

Ce mémoire a présenté le travail effectué dans le cadre d'une compétition Kaggle, comprenant l'analyse des données, la préparation des données, la modélisation et l'évaluation des modèles. Les résultats obtenus démontrent que les modèles sont efficaces, en particulier le SVM qui a été choisi pour la soumission finale. La régression logistique a enregistré un taux de classification satisfaisant de 78,38% avec une AUC de 0.86, tandis que le Random Forest a enregistré une AUC de 0.88, et le SVM a enregistré une AUC de 0.88 avec un taux de classification satisfaisant de 78,8%.

Ce projet a été réalisé en accord avec les objectifs du Master MIASHS, en utilisant les compétences en statistiques et en sciences des données, tout en incorporant l'apprentissage automatique et la visualisation des données. Grâce à cela, il a été possible de saisir et d'analyser un vaste ensemble de données, une compétence recherchée dans la formation.

8.1.2 Impact et perspectives futures

Ce projet a été important en soulignant deux points : il a donné l'opportunité d'appliquer et de renforcer des compétences en science des données et en machine learning, et il a mis en évidence l'importance de la méthodologie et de la rigueur dans la gestion de projet. Il est possible d'améliorer les modèles, d'approfondir les compétences techniques et de participer à de nouvelles compétitions afin de poursuivre l'apprentissage et le progrès dans le domaine de la science des données.

De plus, ce projet a montré l'importance d'être flexible, créatif et d'avoir une réflexion critique dans le domaine de la science des données. Il a également souligné l'importance de saisir comment élaborer un modèle efficace, stable et applicable, en commençant par l'analyse des données et en passant par leur prétraitement. Cette expérience revêt une grande valeur pour une carrière future dans le domaine de la science des données, en soulignant l'importance de la validation externe des modèles prédictifs et de gérer efficacement le temps afin de produire des résultats pertinents dans un contexte professionnel.

Je n'ai pas été impliqué dans une situation d'alternance en entreprise, ce qui m'a épargné la responsabilité sociétale des entreprises (RSE) ou la responsabilité sociale et durable (DDRS). Étant donné le caractère fictif des données du projet en raison de son contexte, il n'y a aucune possibilité d'utilisation de données sensibles ou privées.

En résumé, ce mémoire a illustré comment les compétences acquises lors de ma première année en Master MIASHS ont été mises en pratique, tout en mettant en évidence l'importance de la méthodologie et de la rigueur dans la gestion de projet. Il a également préparé le terrain pour de futures contributions dans le domaine de la science des données.

Bibliographie

[1] Kaggle. Kaggle : Your machine learning and data science community. `https://www.kaggle.com`, 2010.

[2] M. Servajean S. Lèbre, A. Sallaberry. Rentrée m1 miashs septembre 2023, 2023.

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

[4] C. Trottier and M. Amico. Régression logistique et modèles log-linéaires, 2024. Cours de Master MIASHS.

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

[6] Project Jupyter. Jupyter notebook, 2014.

[7] Microsoft Corporation. Microsoft to do, 2017.

[8] Gusthema. Spaceship titanic with tfdf, 2023.

[9] Wikipedia contributors. Pearson correlation coefficient.

[10] Maximilien Servajean. Random forest tp, 2023.

[11] Wes McKinney. Pandas.

[12] Travis E. Oliphant. Numpy.

[13] David Cournapeau. Scikit-learn.

[14] Michael Droettboom John Hunter. Matplotlib.

[15] Michael Waskom. Seaborn.

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

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

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

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

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

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

[22] Marine Demangeot. Classification supervisée et non supervisée - svm, 2022. Cours de Master MIASHS.

[23] Maximilien Servajean. *Arbres de décision et régression - CART*, 2024.