Travail d'étude et de recherche appliquée sur la modélisation et la prédiction de survie des individus face au naufrage du Titanic

Mulapi Tita Ketsia
Master 1 en Science de Données à l'Université de Rouen Normandie
2021-2022
contact@ketsiamulapi.com

1 Résumé

Le naufrage du Titanic a mis fin au paquebot qui devait relier Southampton à New York. Il se déroule dans la nuit du 14(à partir de 22 heure) au 15 avril 1912 dans l'océan Atlantique Nord. (Wikipédia)

Ce travail est principalement axé sur la prédiciton des survies des individus présents au moment des faits. Nous avons eu l'occasion de comprendre les caractéristiques qui font ces individus et, à ressortir les informations cachées permettant de savoir ce qui influence réellement la survie. A cet effet, conscient des tenants et des aboutissants de relever un tel défi, 4 parties parmis les quelles on retrouve, la phase de prétraitemeent des données, la phase de la modélisation, l'explicabilité des algorithmes tout juste avant les résultats, ont été des étapes essentielles à l'aboutissement de ce travail. Nous notons aussi notre approche pédagogique active qui nous a permis d'aller plus loin en découvrant l'univers de l'intelligence artificelle faible, en approchant le concepte de boîte noir des algorithmes et, notre sens de la pédagogie par l'erreur qui revient à démontrer les point faibles et fort d'une intelligence artificielle forte à l'état de l'art.

2 Les données

Disposées en 2 fichiers dont l'un dédié à l'apprentissage et l'autre aux tests, nous avons eu à faire à 11 vairables d'une part, et 10 d'autre part. En effet, dépourvue de la cible (target ou label Survived), la 11e colonne du jeux de test trouve son existence sur l'espace de la compétition kaggle. On retiendra aussi que selon la légende, sur 2224 individus (équipage y compris), environ 1502 ont succombé.

Aussi, en situant le problème d'un point de vue technique, il s'avère que la prédiction d'une valeur entière nous positionne dans un reférentiel de classification, ainsi, tout au long de ce travail on s'intérressera à des modèles de classification à une seule exception : Nous allons inférer sur l'age, de sorte à compléter les valeurs manquantes avec un arbre de décision prédictif. Et donc après avoir rencontré ces problèmes de donnés manquantes et ou abbérantes, nous avons procédé par un ensemble de traitement dit prétraitement qui se résume en :

1. La suppression de la variable Ticket pour son aspect alpha numérique qui d'ailleurs la rend unique et donc inutile à notre égard, la variable nom, qui est assez composé et qui pouvait certe être décomposé mais, en notre sens a été jugé comme n'étant pas nécessaire et, la variable cabine qui était quasi-absentes à plus de 80 pourcents.

2. Encodage des variables Sex et Embarked pour le caractère catégorielle, étant ainsi effectué à l'aide d'un "getdummies" qui applique en arrière plan un algorithme dit one hot encoding fonctionnant sur un principe de binarisation

3. Pour les données d'apprentissages, et uniquement celui-ci, suppression des 2 lignes ou inidividus n'ayant aucune valeur d'embarquation

4. Pour nos données de test, nous avons remplacer par 0, la valeur du prix du billet qui manquait au 152e individu.

Tout ceci juste avant de remplacer la valeur de l'âge, grâce à notre modèle prédictif.

Enfin, notons aussi que de notre exploration, il en ressort des statistiques qui nous ont emmener à réaliser les affirmations suivantes :

1. Plus le tarif de votre ticket est élevé, plus vous avez des chances de survie et donc, appartenir à la première classe sociale, semble privilégier la survie contrairement à la 3e classe

2. Les femmes et les plus jeunes sembles prioritaires dans cette situation

3. Les personnes appartenannt à des regrouppements c'est à dire soit ayant des frères, soeurs ou nounou (SibSp) soit, ayant un nombre de parents et d'enfant élevé (Parch) survivent également

4. Et à notre grande surprise, on constate que même le lieux d'embarquement joue un rôle capital.

3 Les algorithmes utilisés

3.1 Decision Tree

Basé sur le principe CART (Classification And Regression Tree) qui a donnée l'essor à la ségmentation des donnés. Il s'agit d'un algorithme qui fonctionne par supervision et qui s'applique aussi bien dans des cas de classification et de prédiciton. Elle se base sur la notion de ressemblance par les classes, il est compréhensible comme algo. mais fait du surapprentissage, On s'y intérresse car l'algo cherche la variable qui sépare mieux les 2 populations en testant les variables et en utilisant une métrique qu'on appelle le critère. Elaguer les feuilles et les colonnes afin de ne pas avoir un arbre profond permet de ne pas sur apprendre.

3.2 Randome Forest

Il s'gait d'un ensemble d'arbre de décision, le principe est le même, ici lorsqu'on classifie on s'intérresse à la catégorie la plus fréquente alors que dans le cas d'espèce de prédicition, on s'intérresse à la moyenne des valeurs prédites. Ce modèle dit-on, ne déçoit jamais et d'ailleurs dans notre contexte il se distingue aussi entant que meillleur modèle car, toujours performant, même face à des problèmes complexes, on l'utilise en principe lorsqu'on a beaucoup de features.

3.3 XgBoost

Il s'avère que l'on s'est trompé sur les arbres de décision et que le critère n'est pas si efficace qu'on ne le pense, d'où la création du Boosting Gradient, il est supervisé et combine les modèles les plus fort et les plus fiables afin de fournir une meilleur décision. Il peut de ce fait posséder plusieurs possibilité de paramètres, et ce sont ces derniers qui font sa force. Ici aussi on peut limiter la taille de l'arbre pour qu'il évite l'overfitting.

3.4 Gaussian NB

Simple, robuste et effcace mais attention pour GaussianNB tout est indépendant alors que dans notre cas, qui est la réalité, c'est absolument faux, ce qui fait qu'il ne nous permet pas d'avoir de bons résultats au momeent du test, à cause justement de son caractère Naives. Il se base sur le théorème de Bayes.

3.5 Regression Logistique

La regression logistqiue nous permet principalement de faire de la classification binaire. Il s'utilise dans les cas d'une apprentissage supervisé qui prédit la probabilité d'une vairable cible, sa nature, ou de la variable dépendante est dichotomique. Elle se base sur la notion du maximum de vraissemblance qui consiste à traiter le problème comme un problème d'optimisation ou de recherche d'hyperparamètre pour ajuster la probabilité.

3.6 SVM : Support Machine Vector

ici on classe les données en introduisant un hyperplan. Il propose de ce fait des types de modèles linéaire : LinearSVC et SVC tout court ! ils sont légèrmeent différent dans le sens où le premier effectue un One versus All pendant que le second fait un One versus One.

3.7 KNN : K plus proche voisin

Ici l'approche suppose que les familles s'entraident et si un membre en ressort vivant, il y'a de forte chance pour que ceux de son entourage s'en sortent également. Il est donc puissant et efficace, similaire à la regression, à la différence que le KNN possède un paramètre principal qui est le nombre de voisin(s).

4 Explicabilité

4.1 Shape

Il s'gait d'une librairie crée en 2017 pour faire de l'explicabilité.

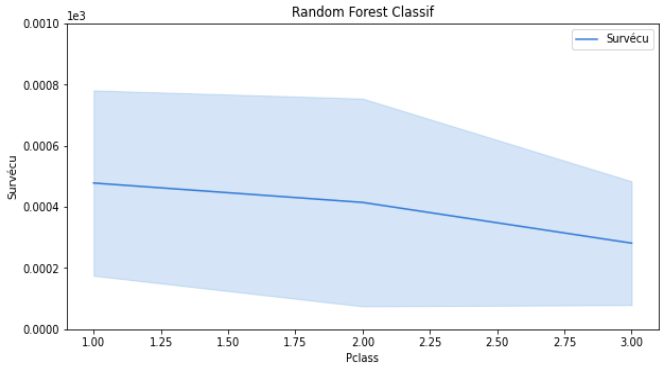
Elle permet de calculer pour chaque features de chaque individu de notre dataset, une importance de ces shapes values. Une fois calculés, on peut en déduire l'importance globale du modèle ou en faire des sous ensembles qui ont du sens.

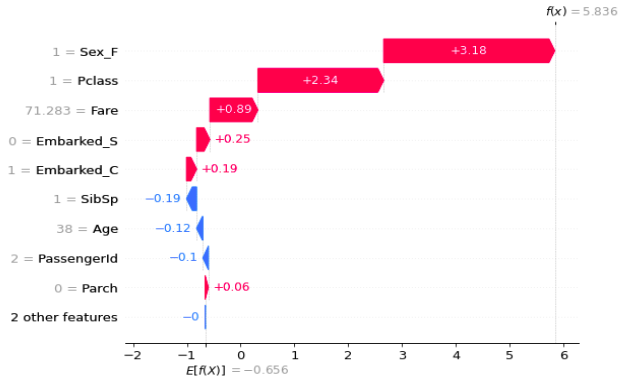
4.2 Skater

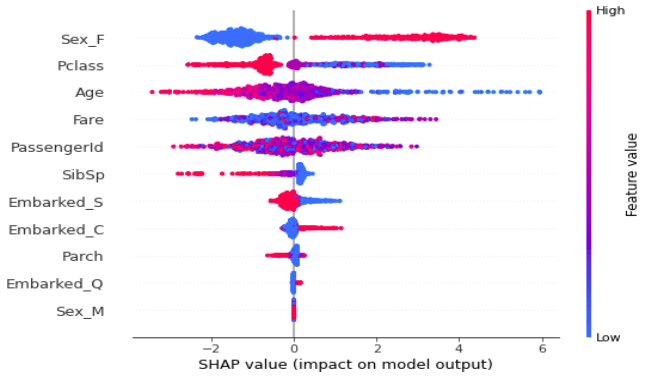
Skater est un cadre unifié permettant l'interprétation de modèles pour toutes les formes de modèles afin d'aider à créer un système d'apprentissage automatique interprétable souvent nécessaire pour les cas d'utilisation du monde réel. Il s'agit d'une bibliothèque python open source conçue pour démystifier les structures apprises d'un modèle de boîte noire à la fois globalement (inférence sur la base d'un ensemble de données complet) et localement (inférence sur une prédiction individuelle).

4.3 Exemples

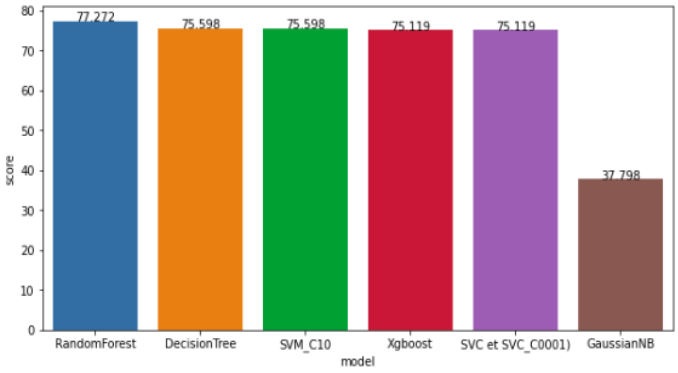
Quelques exemples :







- Test :



References

[1] Compétition Titanic : https://www.kaggle.com/c/titanic/submit.

[2] Skater : https://github.com/oracle/Skater.

[3] Shape : https://github.com/oracle/Skater

[4] Librairie Scikit-Learn

[5] Cours de Théorie Bayésienne de la décision Licence 3 Univeristé de Rouen Normandie . Heutte Laurent

[6] Cours d'Apprentissage Automatique Master 1 Univeristé de Rouen Normandie . Heutte Laurent

[7] Cours de Statistiques des données Master 1 Univeristé de Rouen Normandie . PetitJean Caroline

[8] Cours d'optimisation Master 1 Univeristé de Rouen Normandie . Berar Maxime

5 Résultats

- Apprentissage :