Projet 8 - Participez à une compétition Kaggle !

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Cyril JAUDET, PhD

Physicien médical

Ilyass MOUMMAD

Doctorant

Centre Baclesse

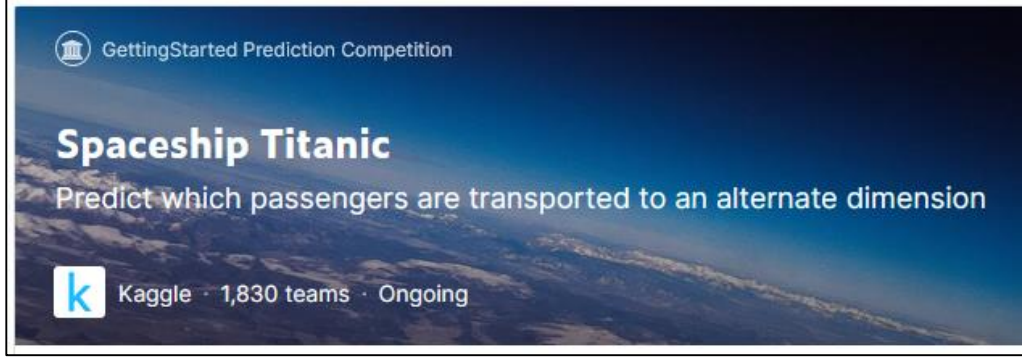
L'excellence pour vaincre votre cancer

Rappel de l’appel à projet

- Contexte :
  - Le site Kaggle propose des compétitions informatiques sur des sujets différents

- Objectif :
  - Participer à une de ces compétitions réelle et en cours
  - Obtenir des résultats mesurables avec un classement
  - Collaborer avec d’autres compétiteurs ou en équipe pour améliorer les modèles
  - Présenter un notebook explicatif de la démarche pour participer à l’évolution collective

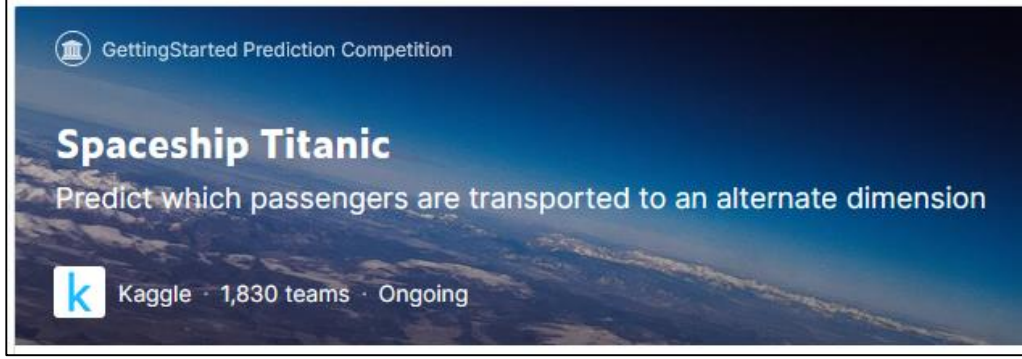
Compétition choisie



- Contexte :
  - Un navire spatial comprenant 13 000 passagers a traversé une anomalie spatio-temporelle
  - L'ordinateur de bord d'un précédent voyage ayant connu la même fin nous informe de données de passagers ayant ou non était transporté dans la faille spatio-temporelle
- Objectif :
  - A partir des informations de bord (nom des passagers, n° de cabin…), être capable de prédire avec des modèles de classification de machine learning ou de deep learning, si les passagers peuvent être sauvés ou non

URL de la compétition : https://www.kaggle.com/competitions/spaceship-titanic/overview

Compétition choisie



File and Data Field Descriptions

- **train.csv** - Personal records for about two-thirds (~8700) of the passengers, to be used as training data.
  - `PassengerId` - A unique Id for each passenger. Each Id takes the form `gggg_pp` where `gggg` indicates a group the passenger is travelling with and `pp` is their number within the group. People in a group are often family members, but not always.
  - `HomePlanet` - The planet the passenger departed from, typically their planet of permanent residence.
  - `CryoSleep` - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
  - `Cabin` - The cabin number where the passenger is staying. Takes the form `deck/num/side`, where `side` can be either `P` for *Port* or `S` for *Starboard*.
  - `Destination` - The planet the passenger will be debarking to.
  - `Age` - The age of the passenger.
  - `VIP` - Whether the passenger has paid for special VIP service during the voyage.
  - `RoomService`, `FoodCourt`, `ShoppingMall`, `Spa`, `VRDeck` - Amount the passenger has billed at each of the *Spaceship Titanic*'s many luxury amenities.
  - `Name` - The first and last names of the passenger.
  - `Transported` - Whether the passenger was transported to another dimension. This is the target, the column you are trying to predict.
- **test.csv** - Personal records for the remaining one-third (~4300) of the passengers, to be used as test data. Your task is to predict the value of `Transported` for the passengers in this set.
- **sample_submission.csv** - A submission file in the correct format.
  - `PassengerId` - Id for each passenger in the test set.
  - `Transported` - The target. For each passenger, predict either `True` or `False`.

Centre Baclesse
L'excellence pour vaincre votre cancer

Nettoyage des données

- Pas de valeurs aberrantes
- Gestion des NaN :
  - Planète d'origine / Destination : mettre la planète la plus fréquente *(Europa* et *Trappist-1e)*
  - Dépense totale : si cryosleep = True alors mettre 0 sinon mettre la valeur moyenne
  - VIP : mettre à *False* car sont les non VIP sont très majoritaire
  - Side/Deck : mettre de façon aléatoire une lettre car les proportions sont homogènes
  - Cabin number : mettre un chiffre aléatoire entre 1 et 1894
  - Cryosleep : mettre la situation la plus fréquente *(False)*

Features engineering

- Informations crées :
  - *« Dépenses totales »*
    - En utilisant les variables de frais de service utilisés
  - *« FirstName » et « LastName »*
    - En utilisant la variable “*Name" (supposant qu’une famille à plus de probabilité de rester ensemble)*
  - *« Deck », « Side » et « Cabin number »*
    - En utilisant la variable “*Cabin" (car celle-ci regroupait plusieurs informations)*

Centre Baclesse
L'excellence pour vaincre votre cancer

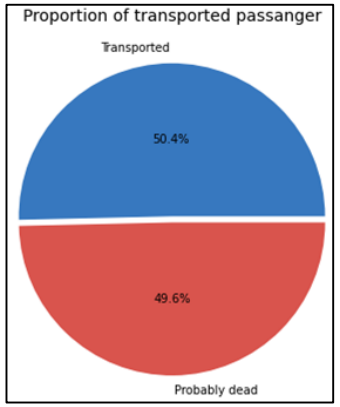
Modélisation

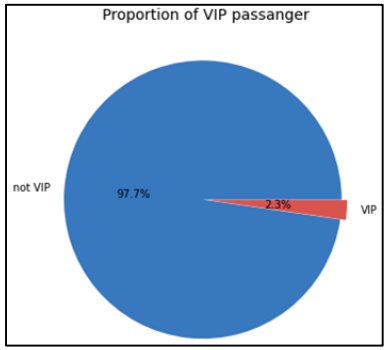
- Features utilisés pour l'entrainement des modèles :
  - Features **quantitatifs** :
    - Age
    - Cabin_number
    - Dépense_totale
    - RoomService
    - FoodCourt
    - ShoppingMall
    - Spa
    - VRDeck
  - Features **catégoriels** :
    - Firstname
    - Lastname
    - Home Planet
    - Destination Planet
    - VIP
    - Cryosleep
    - Deck
    - Side

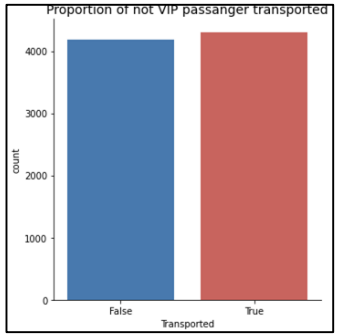
Utilisation d'un OneHotEncoder pour les variables catégorielles

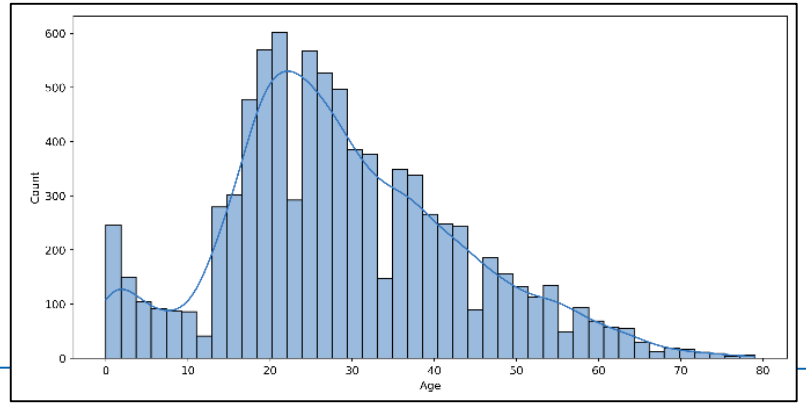
Exploration des données

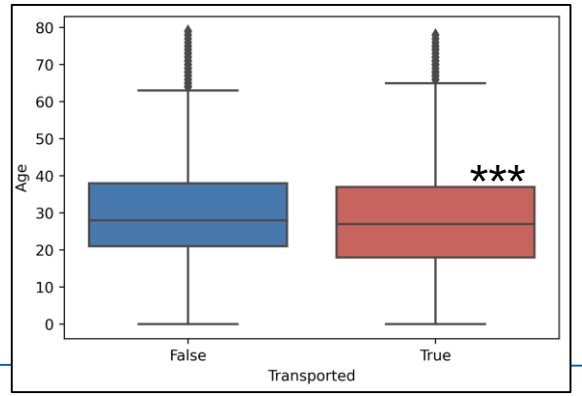
- Proportion des passagers transportés
  - Autant de passager dans une classe que dans une autre
- VIP or not ?
  - Très peu de VIP et cela n'a pas d'importance
- Age des passagers
  - Distribution assez homogène dans les adultes
  - Les plus jeunes ont cependant plus de chance de survivre









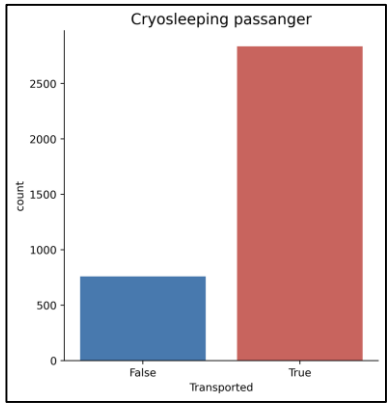


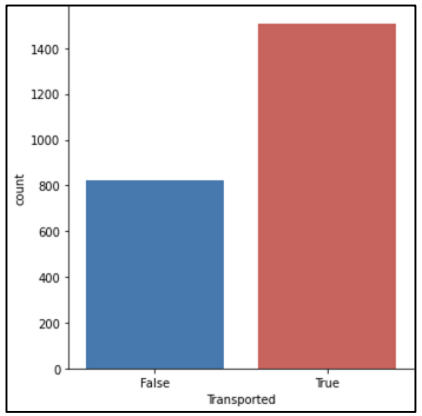
***p<0.001

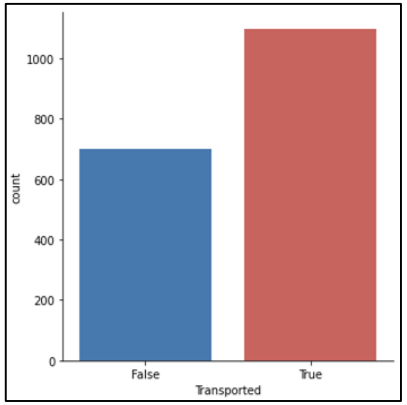
Centre Baclesse
L'excellence pour vaincre votre cancer

Exploration des données

- Critère sommeil cryogénique
  - Les personnes en sommeil cryogénique ont beaucoup plus survécu
- La planète d'origine a-t-elle un impact ?
  - Les passagers venant d'*Europe* semblent avoir été plus chanceux
- La destination a-t-elle un impact ?
  - Les passagers pour *55 Cancri e* semblent avoir été plus chanceux







Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation

- Choix des modèles de classification étudiés :

Machine Learning :

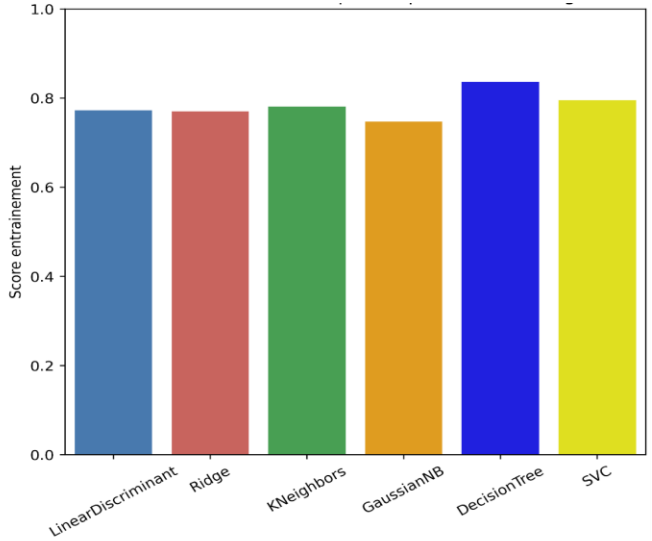
- LinearDiscriminant()
- Ridge()
- CatBoost()
- KNeighbors()
- GaussianNB()
- RandomForestClassifier()
- SVC()

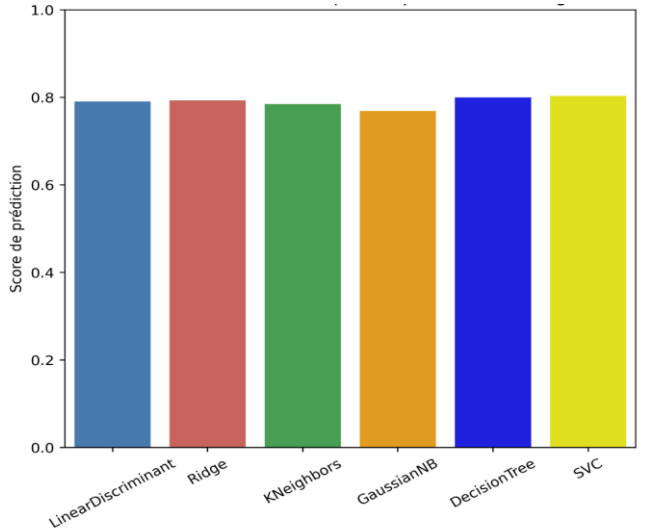
Deep Learning :

- **Tensorflow()**
- **Pytorch()**

Modélisation Machine Learning

- Modèles testés :
  - LinearDiscriminant, Ridge, KNeighbors, GaussianNB, RandForestClassifier, SVC
- Optimisation des hyperparamètres par Gridsearch et validation croisée
- Evaluation des performances par études du score d'entrainement et de prédiction



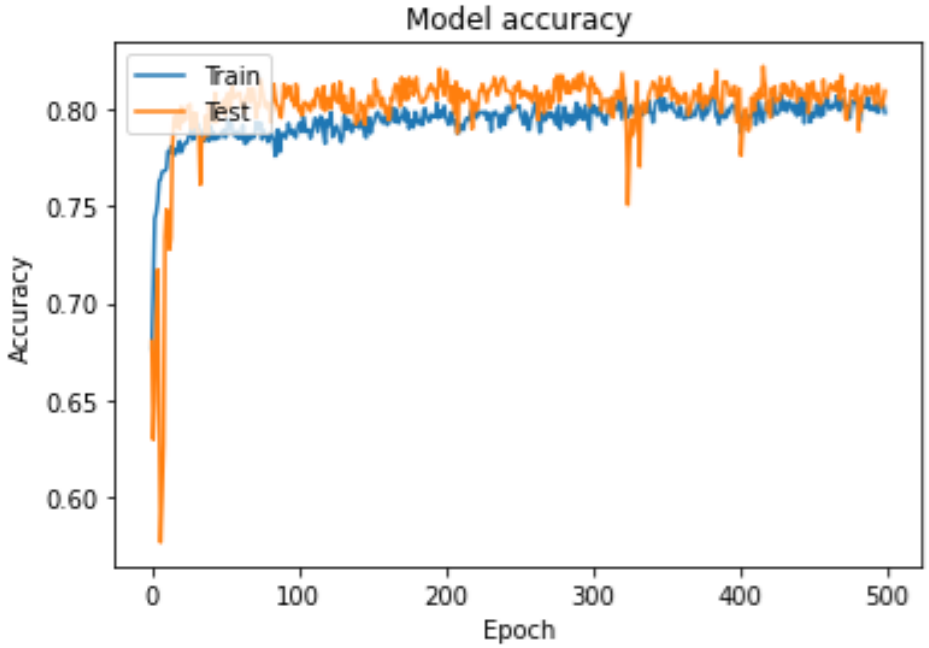


Les modèles RandomForestClassifier et SVC donnent les meilleurs résultats (0.82)

Centre Baclesse
L'excellence pour vaincre votre cancer

Modélisation Deep Learning n°1

- Modèle utlisé :
  - Tensorflow.Keras
- Architecture :
  - 1 couche de neurone d'entrée (relu, 30 features)
  - 2 couches de neurones cachées (relu, 782 neurones)
  - 1 couche de sortie (softmax, 2 output de prédiction)
- Paramètres d'optimisation :
  - **optimizer** = "*Nadam*" ;
  - **loss** = "*BinaryCossentropy*" ;
  - **metrics** = "*BinaryAccuracy*"
  - 641 242 paramètres entrainés



Centre Baclesse
L'excellence pour vaincre votre cancer

Le modèle de DL-01 avec des paramètres choisis nous donne une accuracy de 0.82

Modélisation Deep Learning n°2

- Modèle utlisé :
  - Tensorflow.Keras + tunning des hyperparamètres avec ***keras-tuner***
- Hyperparamètres testés :
  - **Activation couches cachées :** "*elu, gelu, relu, selu*"
  - **Activation couche sortie :** "*sigmoid, hard_sigmoid, softmax, swish, tanh*"
  - **Range nb neurone** : "100=>len(X)/10" ;
  - **Learning_rate :** "0.0005 à 0.1"
- Meilleurs hyperparamètres :
  - **300 et 500** neurones avec **relu, sigmoid** en sortie**,** learning rate = **0.05**

Centre Baclesse
L'excellence pour vaincre votre cancer

Le modèle de DL-02 avec des paramètres choisis nous donne une accuracy de 0.80

Modélisation Deep Learning n°3

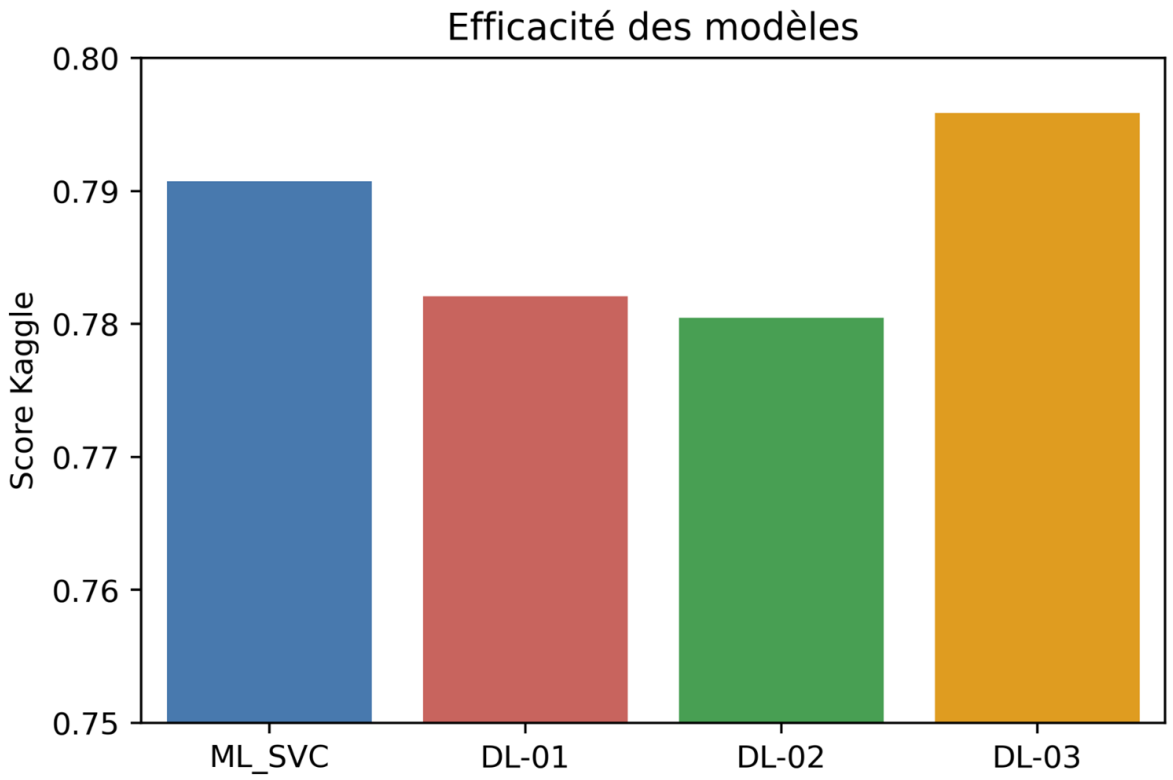
- Modèle utlisé :
  - Pytorch + MultiLayerPerceptron
- Architecture :
  - 1 couche de neurone d'entrée (relu, 30 features)
  - 3 couches de neurones cachées (relu, 256 neurones, dropout=0.2)
  - 1 couche de sortie (relu, 1 valeur de prédiction)
- Paramètres d'optimisation :
  - **optimizer** = "*Adam*" ;
  - **loss** = "*BCEWithLogitsLoss*" ;
  - **metrics** = "*BinaryAccuracy*"
  - 135 809 paramètres entrainés

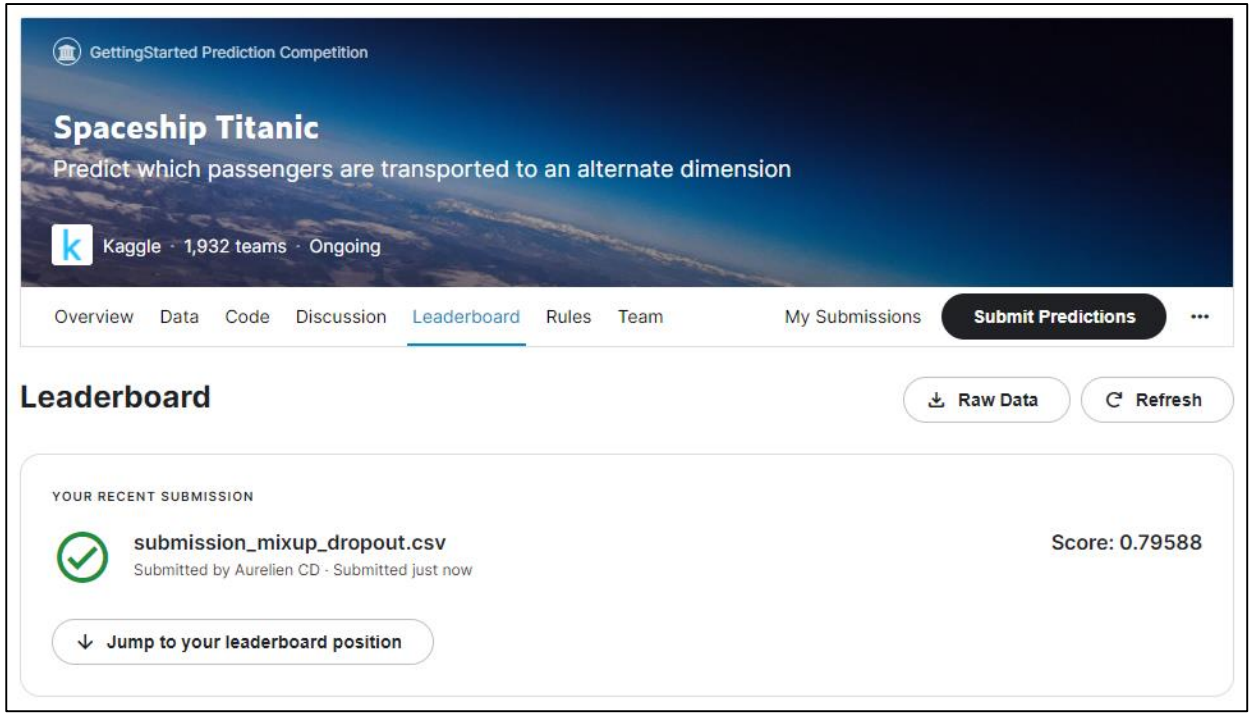
Centre Baclesse
L'excellence pour vaincre votre cancer

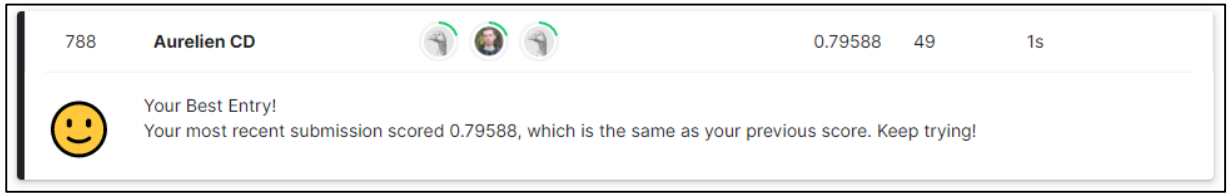
Le modèle de DL-03 avec des paramètres choisis nous donne une accuracy de 0.82

Prédiction et performance des modèles

- Score dans la compétition :







Centre Baclesse
L'excellence pour vaincre votre cancer

Notre classement pour cette compétition est : 788/1932

Présenter un notebook explicatif / participer à l'évolution collective

- Travail en équipe :
  - Cyril JAUDET, PhD, Physicien médical / Ilyass MOUMMAD, Doctorant
  - Utilisation de GitHub pour suivi du code (utilisation de branch)
    - https://github.com/AurelienCD/Formation_OCR_Ing_Machine_Learning/blob/main/P8_01_notebook.ipynb
- Kernel Kaggle :
  - Kernel de présentation de notre approche dans un notebook :
    - https://www.kaggle.com/aureliencd/ocr-projet-8-iml-acdcjim
- Discussion avec d'autres participants :
  - Test et discussion d'une approche d'un autre kernel
    - https://www.kaggle.com/code/sardorabdirayimov/awesome-nn-titanic-disaster/notebook?scriptVersionId=90946453

Centre Baclesse
L'excellence pour vaincre votre cancer

CONCLUSION

- Rappel de la problématique :
  - Participer à une compétition Kaggle
  - Travailler en équipe ou discuter avec d'autres participants

- Résultats :
  - Le modèle de ML avec SVC() propose de bons résultats
  - Cependant le modèle de DL utilisant Pytorch nous donne les meilleurs résultats avec un score de 0.79588
  - Notre place dans cette compétition est 788ème sur 1982 (à la date du 04 mai 2022)

Centre Baclesse
L'excellence pour vaincre votre cancer

Projet 8 - Participez à une compétition Kaggle !

Aurélien Corroyer-Dulmont, PhD

Ingénieur imagerie médicale

Cyril JAUDET, PhD

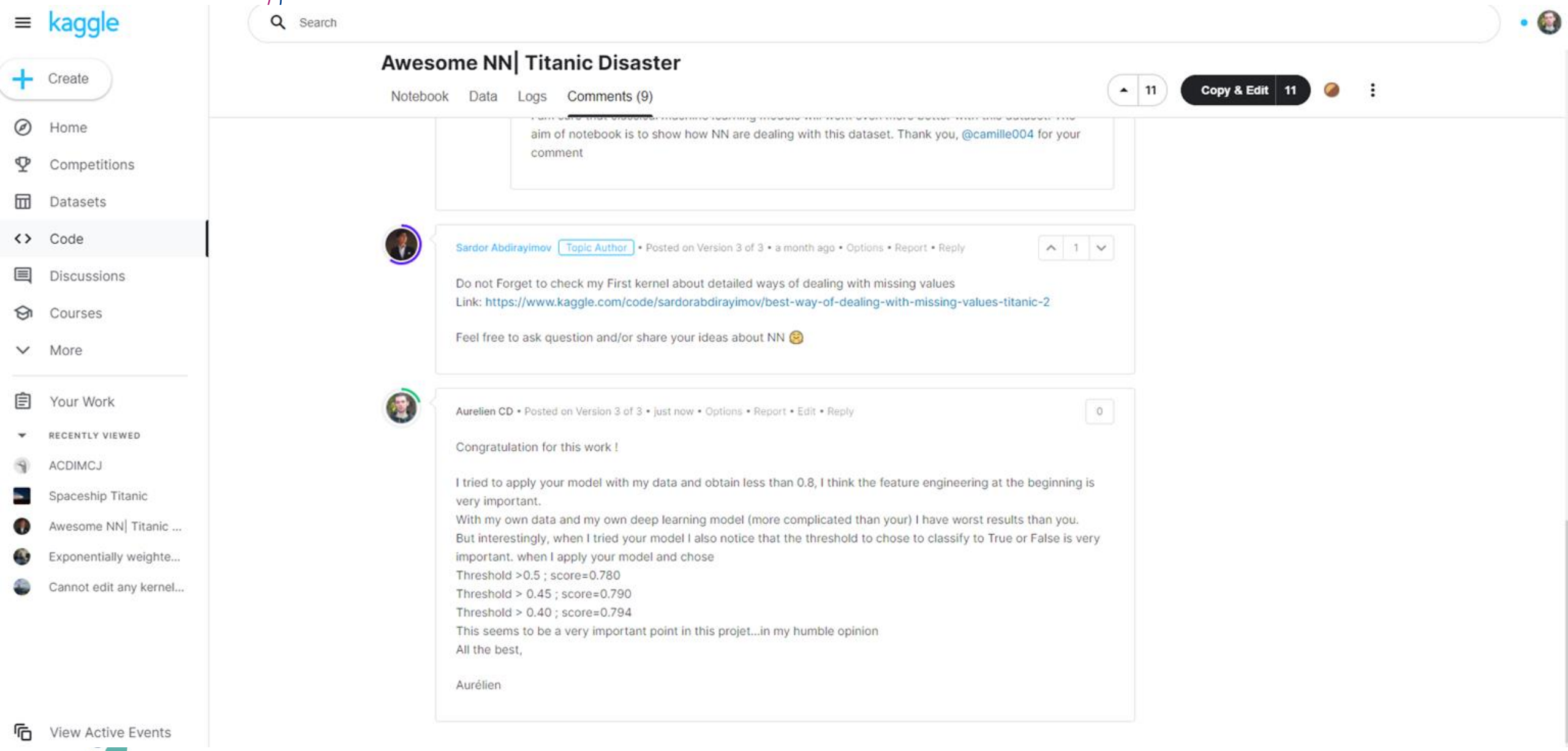
Physicien médical

Ilyass MOUMMAD

Doctorant

Centre Baclesse

L'excellence pour vaincre votre cancer


kaggle
Search
Create
Home
Competitions
Datasets
Code
Discussions
Courses
More
Your Work
RECENTLY VIEWED
ACDIMCJ
Spaceship Titanic
Awesome NN| Titanic ...
Exponentially weighte...
Cannot edit any kernel...
View Active Events
Awesome NN| Titanic Disaster
Notebook
Data
Logs
Comments (9)
11
Copy & Edit
11
aim of notebook is to show how NN are dealing with this dataset. Thank you, @camille004 for your comment
Sardor Abdirayimov
Topic Author
• Posted on Version 3 of 3 • a month ago • Options • Report • Reply
1
Do not Forget to check my First kernel about detailed ways of dealing with missing values
Link: https://www.kaggle.com/code/sardorabdirayimov/best-way-of-dealing-with-missing-values-titanic-2
Feel free to ask question and/or share your ideas about NN
Aurelien CD • Posted on Version 3 of 3 • just now • Options • Report • Edit • Reply
0
Congratulation for this work !
I tried to apply your model with my data and obtain less than 0.8, I think the feature engineering at the beginning is very important.
With my own data and my own deep learning model (more complicated than your) I have worst results than you.
But interestingly, when I tried your model I also notice that the threshold to chose to classify to True or False is very important. when I apply your model and chose
Threshold >0.5 ; score=0.780
Threshold > 0.45 ; score=0.790
Threshold > 0.40 ; score=0.794
This seems to be a very important point in this projet...in my humble opinion
All the best,
Aurélien