# ResponsibleML based models for efficient allocating liver in transplantation process in the US

Hoang Thien Ly

May 22, 2022

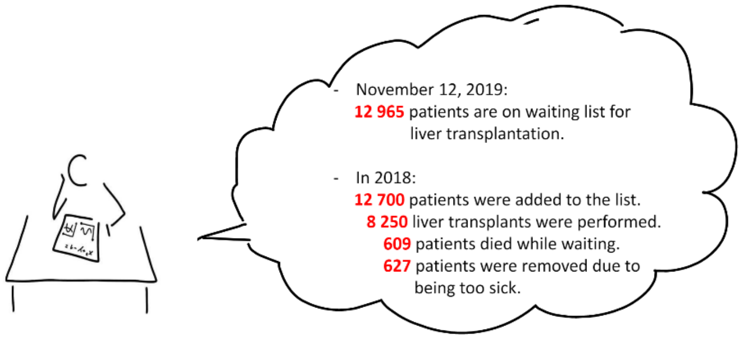
**Abstract**

The first successful liver transplant was performed on May 5, 1963. Thenceforth, liver transplantation for end-stage liver patients has gained worldwide acceptance as an established treatment saving thousands of lives annually. In the scope of the project, we will place our emphasis on the system of allocation livers for end-stage liver disease patients in the US. They used Model for End-stage Liver Disease (MELD-Na) Score to determine how urgently patients need a transplant in order to rank them on a waiting list. Some patients may be able to receive a donor's liver after a few weeks, but in some cases, the waiting time may be up to months or years.

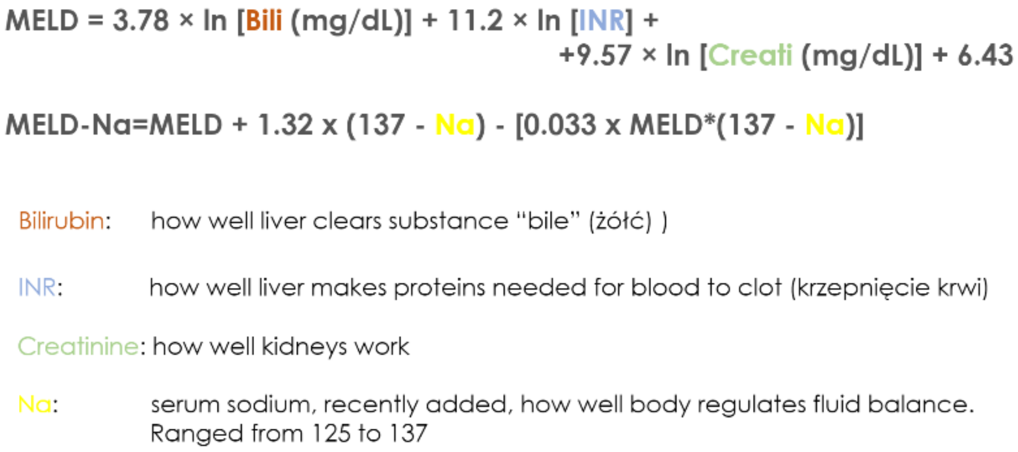
While the MELD is indeed predictive of mortality on the transplant list, the score was fit with only three features. This motivates us to further investigate the MELD-Score and assessing the quality of predictive performance of MELD-score compared with the use of Machine Learning techniques. Furthermore, we will also demonstrate the extendibility of the topic in assessing the fairness of the model toward demographic features such as gender or race, ...etc to overcome the possible gaming due to manipulability by doctors.

## 1 Introduction

From May 5, 1963, we have been witnessing series of fruition about liver transplantation. Simultaneously, t has generated a discrepancy between supply and demand, thereby, generating a persistent insufficiency that results in thousands of candidate deaths annually while awaiting on the waiting list. Given the shortage of donors' livers, one of the most crucial challenges is how to accurately prioritizing sickest patients to be on top of waiting list, so that the limited supply of donated livers can be allocated to maximize the benefit from transplantation.


- November 12, 2019:
12 965 patients are on waiting list for liver transplantation.
- In 2018:
12 700 patients were added to the list.
8 250 liver transplants were performed.
609 patients died while waiting.
627 patients were removed due to being too sick.

Since 2002, the States initially employed the Model for End-Stage Liver Disease (MELD) score to rank severity of patients and, consequently, priority for receiving transplantation. Thanks to the use of MELD-Score, the instant fruition came to hopitalization system in the US. “The MELD-based allocation system was immediately successful, leading to first ever reduction in the number of waiting list candidates and a 15 % reduction of mortality among those on waiting list”, mentioned in Freeman, R., Wiesner, R., Edwards, E., Harper, A., Merion, R., Wolfe, R.: Results of the first year of the new liver allocation plan. Liver Transplant, 10, 7-15 (2004). But specific cohort of patient populations, however, are at risk of death or due to disease progression, become unsuitable for transplantation and removed from the waiting list. The main cause locates at the point, the disease progression is not captured in their lab-based MELD score calculation. To allow better assessing severity of patients, the MELD exception point granting was adopted as a significant weakness in the allocation process, leading to the possible gaming from manipulability by doctors. Notwithstanding, in reality, there remains a higher risk of waitlist death/removal for candidates without exception points, when compared to those with exception points.



$$\mathrm{MELD} = 3.78 \times \ln[\mathrm{Bili}\ (\mathrm{mg/dL})] + 11.2 \times \ln[\mathrm{INR}] + 9.57 \times \ln[\mathrm{Creati}\ (\mathrm{mg/dL})] + 6.43$$

$$\mathrm{MELD\text{-}Na} = \mathrm{MELD} + 1.32 \times (137 - \mathrm{Na}) - [0.033 \times \mathrm{MELD}^{*}(137 - \mathrm{Na})]$$

Bilirubin: how well liver clears substance “bile” (żółć) )

INR: how well liver makes proteins needed for blood to clot (krzepnięcie krwi)

Creatinine: how well kidneys work

Na: serum sodium, recently added, how well body regulates fluid balance.
Ranged from 125 to 137

Figure 1: MELD Score formula

In the scope of the project, we seek to utilize the state-of-the-art machine learning methods combining with Explainable AI methods, to propose an usable ML-based model to tackle those aforementioned issues in the meaning of generating a more accurate prediction of status of patients that would in-return allow better prioritization of candidates awaiting liver transplantation.

# 2 Materials and Methods

## 2.1 Data

**Dataset:** Waitlist, transplant, and follow-up information was obtained from January 1, 2002 to September 5, 2021, from the Organ Procurement and Transplantation Network (OPTN) Standard Transplant Analysis and Research (STAR) dataset.

Dataset contains 329468 observations and 426 features.

## 2.2 Prediction methods

The prediction problem is based on using ML models trained on historical data. Specifically, problem type is defined as classification, where 0 stands for patients that are removed from waiting list due to too sick/ death/ died on transplantation process/ patient died during living donor Transplant procedure and 1 stands for successful transplantation.

**Metric:** performance was also evaluated by out-of-sample area under the curve on the testing set. A model's AUC corresponded to the probability that a randomly drawn observation whose dependent value was 1 (patients with successful transplant) had a higher score under the model than a randomly drawn observations with dependent value was 0.

## 2.3 Tools

Partial Dependence Profile, Break Down method, Permutation-based variable important and Shapley values will be calculated to provide explanations for first-cycle of training models. Technical requirements:

1. Programming language: R
2. Necessary packages: xgboost, LightGBM, catboost, ranger, caret, forester, DALEX (for explaining models) and fairmodels (arXiv:2104.00507).
3. Model training by self-developed **forester** package
4. Member: Hoang Thien Ly
5. Report in LaTEX + Presentation slides

# 3 Results

## 3.1 Methodology

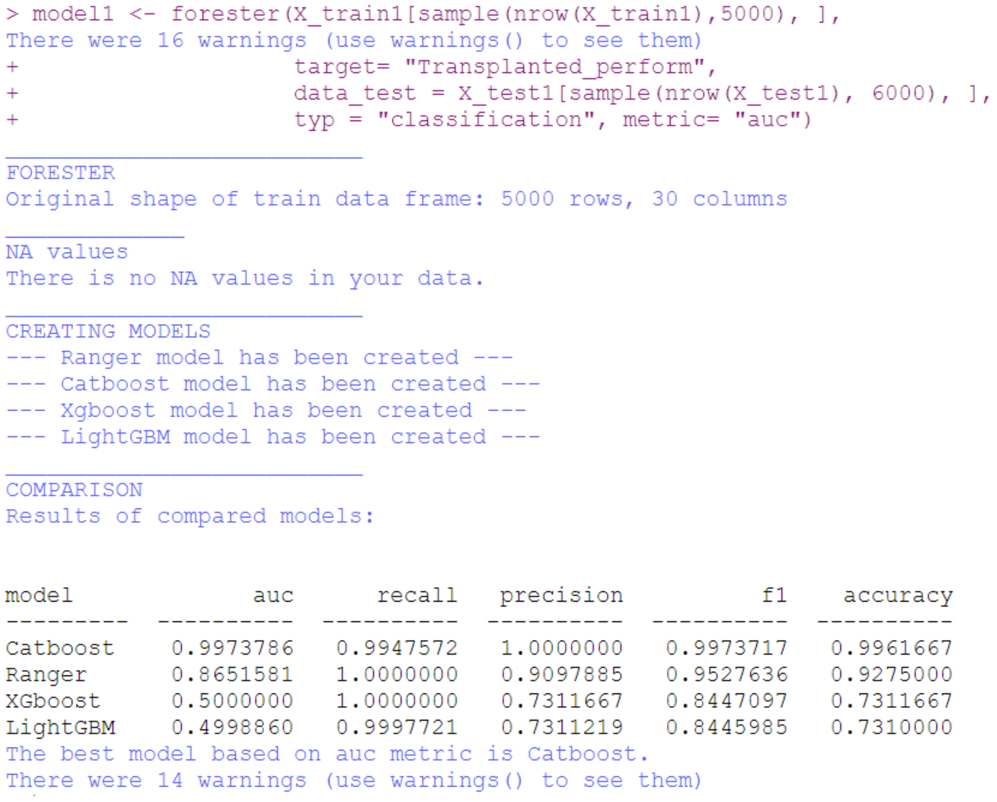
Below, we show the results of models (there are four trained models available in **forester**: random forest, XGBoost, LightGBM and Catboost).

We split the original data frame into training set and test set in the ratio 3:2. Due to the limitation of computing ability, we used only 5000 observations to train the models by **forester**, and 6000 observations to test models' performance.

The idea is, one dataset, we excluded all demographic features and using only laboratory test indices, to train and assess models. Second dataset, we include demographic features such as gender, BMI, education,... After all, the pertubation-based feature important method will be employed on the best returning score model based on AUC, to determine, do the important features proposed by ML models satisfy with domain knowledge of MELD-score formula. And whether some information of demography may improve models' performance? The final step, which is initiative in our project, is to assess the fairness of created model in respect with gender.

## 3.2 Models' performance

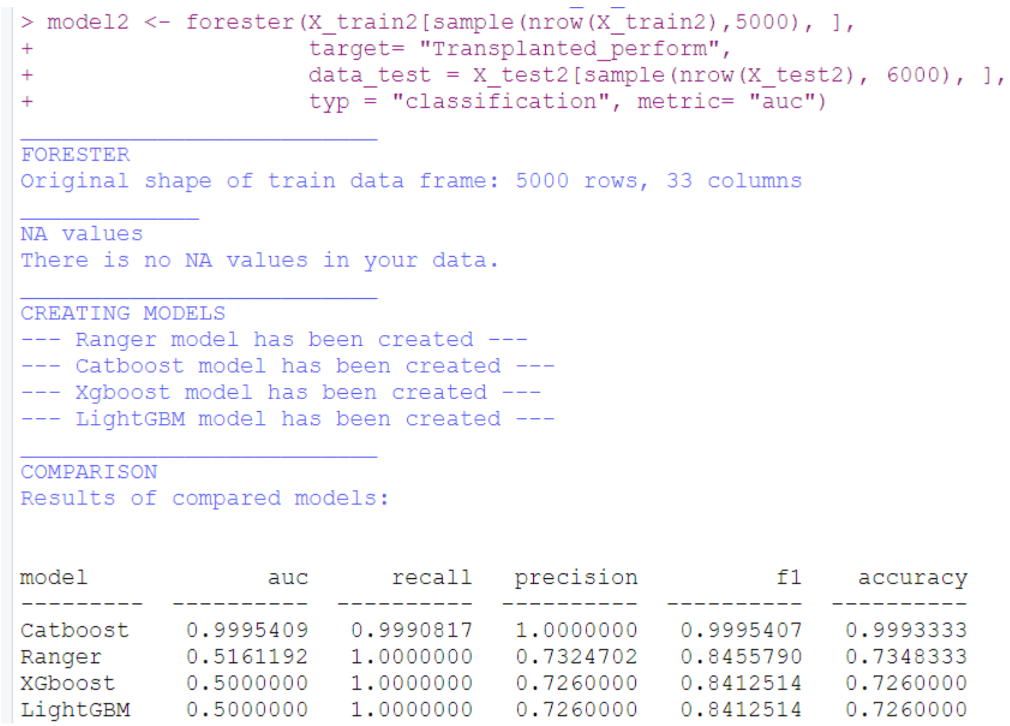
### 3.2.1 On data without demographic features:



```
> model1 <- forester(X_train1[sample(nrow(X_train1),5000), ],
There were 16 warnings (use warnings() to see them)
+                     target= "Transplanted_perform",
+                     data_test = X_test1[sample(nrow(X_test1), 6000), ],
+                     typ = "classification", metric= "auc")

__________________________
FORESTER
Original shape of train data frame: 5000 rows, 30 columns

_____________
NA values
There is no NA values in your data.

__________________________
CREATING MODELS
--- Ranger model has been created ---
--- Catboost model has been created ---
--- Xgboost model has been created ---
--- LightGBM model has been created ---

__________________________
COMPARISON
Results of compared models:


model               auc      recall   precision          f1    accuracy
---------  ----------  ----------  ----------  ----------  ----------
Catboost    0.9973786   0.9947572   1.0000000   0.9973717   0.9961667
Ranger      0.8651581   1.0000000   0.9097885   0.9527636   0.9275000
XGboost     0.5000000   1.0000000   0.7311667   0.8447097   0.7311667
LightGBM    0.4998860   0.9997721   0.7311219   0.8445985   0.7310000
The best model based on auc metric is Catboost.
There were 14 warnings (use warnings() to see them)
```

Figure 2: Models' performance on non demographic dataset

### 3.2.2 On data with demographic features:



```
> model2 <- forester(X_train2[sample(nrow(X_train2),5000), ],
+                     target= "Transplanted_perform",
+                     data_test = X_test2[sample(nrow(X_test2), 6000), ],
+                     typ = "classification", metric= "auc")
__________________________
FORESTER
Original shape of train data frame: 5000 rows, 33 columns
_____________
NA values
There is no NA values in your data.
__________________________
CREATING MODELS
--- Ranger model has been created ---
--- Catboost model has been created ---
--- Xgboost model has been created ---
--- LightGBM model has been created ---
__________________________
COMPARISON
Results of compared models:


model            auc      recall   precision          f1    accuracy
---------  ----------  ----------  ----------  ----------  ----------
Catboost    0.9995409   0.9990817   1.0000000   0.9995407   0.9993333
Ranger      0.5161192   1.0000000   0.7324702   0.8455790   0.7348333
XGboost     0.5000000   1.0000000   0.7260000   0.8412514   0.7260000
LightGBM    0.5000000   1.0000000   0.7260000   0.8412514   0.7260000
```

Figure 3: Models' performance on non demographic dataset

### 3.2.3 Recap part 1

1. Model of Catboost dominates the performance on AUC metric in both cases.
2. In Ranger, performance of model noticeably deteriorates on the data set with demographic features.
3. There is not much difference in AUC metric between dataset with/without demographic features.

## 3.3 Hypothesis 1

We will check, what are the important features proposed by our ML models, and what are currently used in MELD-Score. Then to check, whether we should propose some new components for MELD-Score formula, or giving some supports to omit some lab-based indices.

The method we use is permutation-based variable importance.

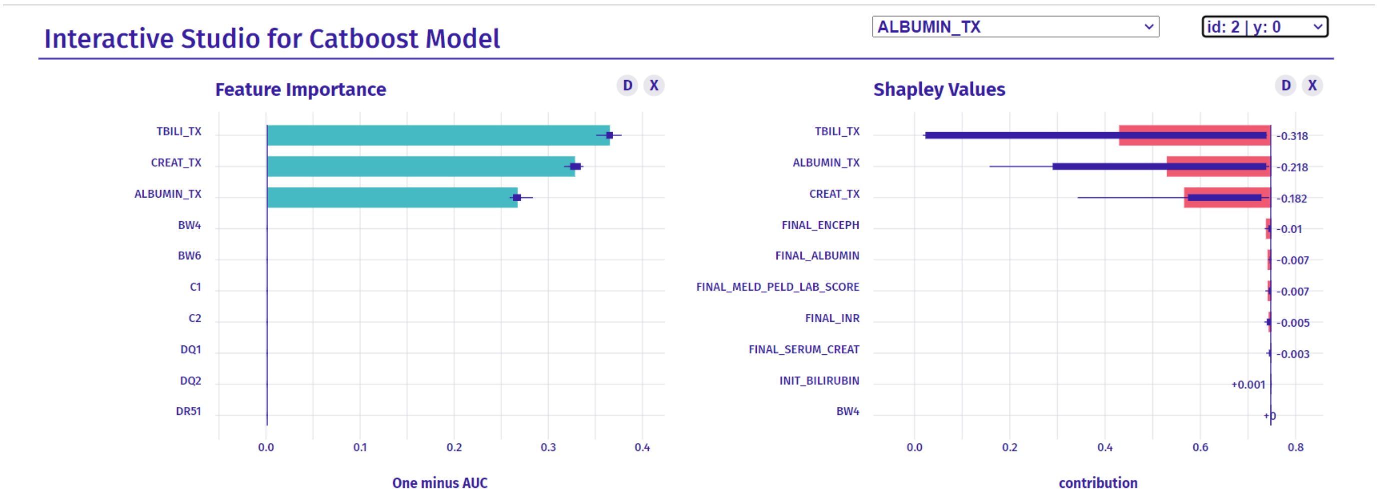


Figure 4: Important features on non-demographic dataset

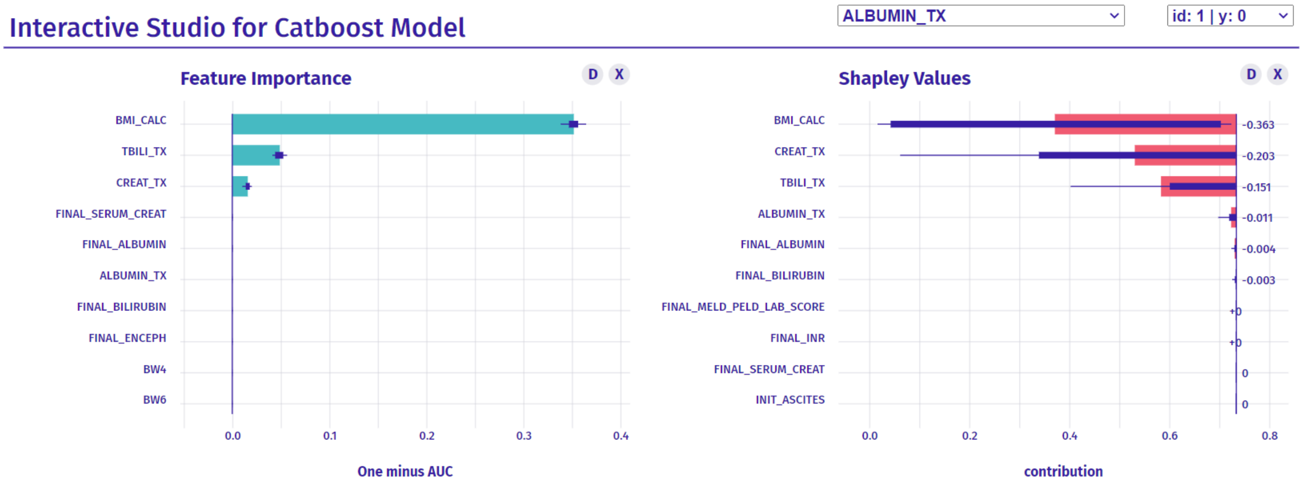


Figure 5: Important features on demographic dataset

In the first section, our model states that three features are most important: “Bilirubin”, “Creatinine” and “Albumin” out of more that 30 other features. Two of them are currently used in MELD-Score. So that, we suggest a reconsideration of adding “Albumin” to the current model of MELD-Score and recalculate the importance of “INR” in our formula.

In the second section, after adding two demographic features: BMI  Gender, our model says, BMI is also significant and should be thoroughly considered to take into account of MELD-Score model.

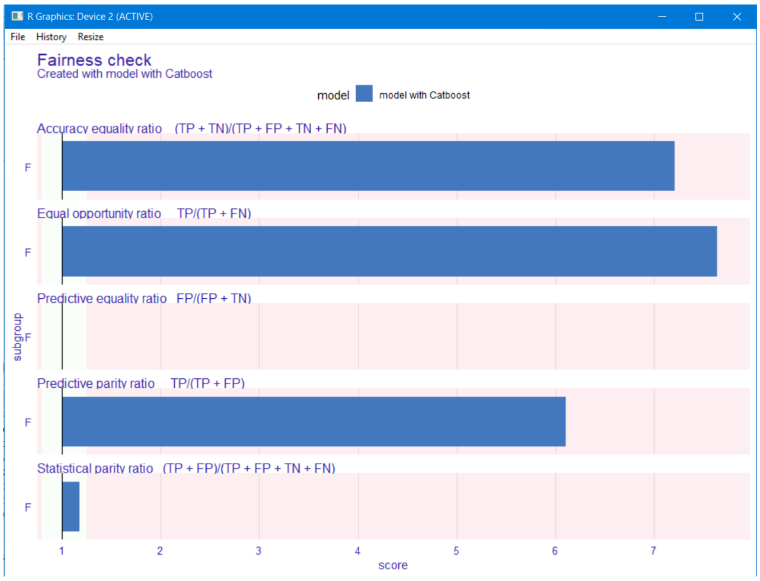


Figure 6: Fairness on created model

# 4 Checking fairness of created model:

Our created model is strongly based for the group of gender Male. The blue bar reached the red region of Male according to some metrics for fairness.

# 5 Summary

Our responsibleML based models show a great potential in predicting mortality of transplant patients with end-stage liver diseases. As well as, it consolidates for currently used features in MELD formula, as well as, proposing the "BMI" should be employed in the MELD-Score (advancement!). But the problem is, our model is biased according to Male group.

The limitation is, if we train model on greater sample size, the model performance could be better.

# 6 References

- Bertsimas, D., Kung, J., Trichakis, N., Wang, Y., Hirose, R., Vagefi, P.: Development and validation of an optimized prediction of mortality for candidates awaiting liver transplantation.Am. J. Transplantat. 19, 1109–1118 (2018)
- Byrd J., Balakrishnan S., Jiang X., Lipton Z.C. (2021): Predicting Mortality in Liver Transplant Candidates. In Shaban-Nejad A., Michalowski M., Buckeridge D.L. (eds) Explainable AI in Healthcare and Medicine. Studies in Computational Intelligence, vol 914. Springer, Cham.
- Waiting List, MELD Score, Liver Transplants — Dr. Robert S.Brown

- Liver Transplant Waitlist, Part 1— UCLA Transplantation Services
- UNOS.org