Modeling Analysis Report on Survival of Allogeneic Hematopoietic Cell Transplantation (HCT) Patients

1. Introduction

Survival analysis is a crucial branch of statistics that focuses on analyzing the time until an event of interest occurs. It is particularly important in medical research, such as evaluating the survival time of cancer patients or the time to complications after transplantation. Survival datasets typically contain censored observations, especially right-censored data, meaning that the event of interest (e.g., death) has not occurred by the end of the study for some individuals, making their exact survival times unknown.

This study utilizes a dataset of patients who underwent allogeneic hematopoietic cell transplantation (HCT). The objective is to evaluate and compare the predictive performance (C-index) of various models and identify key factors that influence patient survival using feature importance analysis.

2. Dataset Overview and Preprocessing

2.1 Dataset Description

1. The dataset comprises 28,800 samples and 60 variables, including:

- 2 target variables: `efs` (event occurrence) and `efs_time` (event time),
- 1 unique patient ID,
- 57 feature variables.

2. Only 7 features are completely free of missing values. 8 features have more than 20% missing data, with the highest missing rate reaching 58%. Most features with fewer than 20 unique values tend to be categorical, which requires proper encoding methods.

3. Regarding censoring, 46% (13,268) of patients are right-censored, indicating they were still alive at the time of data cutoff. The distribution of censored data is skewed to the right. Thus, modeling must appropriately handle censored cases.

2.2 Preprocessing Steps

1. **Feature Removal:** Drop features with missing values exceeding 20%.
2. **Missing Value Imputation:** Fill remaining missing values with `NA` to facilitate model compatibility.
3. **Categorical Encoding:** Encode features with fewer than 20 unique values as Categorical features to accommodate model training.

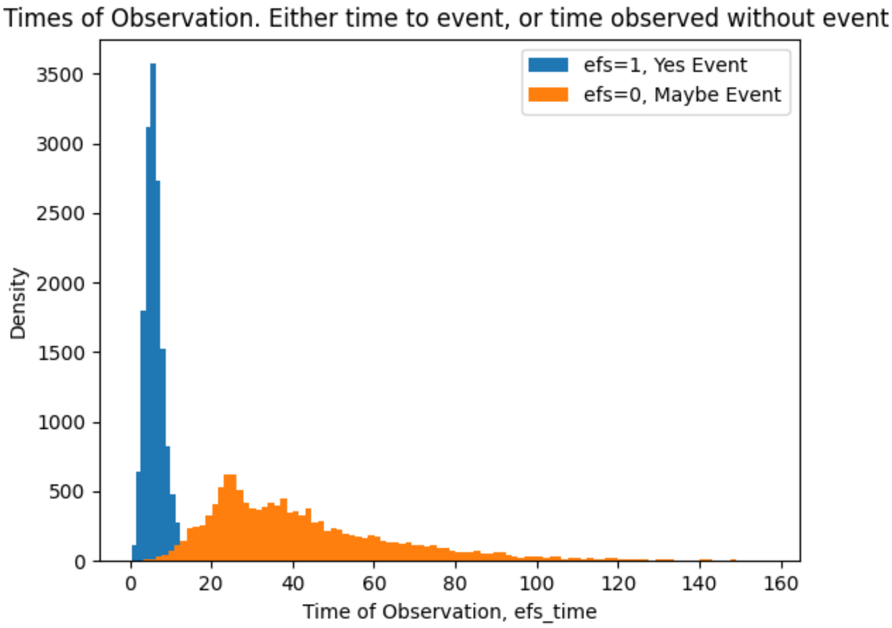


Figure 1: Distribution Plot of Censored Events

3. Modeling Approaches

The primary goal is to predict survival probability based on `efs` and `efs_time`, while accounting for right-censoring. Two baseline survival models are used: the Kaplan-Meier estimator and Cox proportional hazards model. These are then integrated with XGBoost, CatBoost, and LightGBM to construct five hybrid models.

3.1 Kaplan-Meier Estimator

The Kaplan-Meier (KM) estimator is a non-parametric method for estimating survival probability:

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of events at time t_i, and n_i is the number of individuals at risk just before t_i.
In the context of this dataset:

1. Sort the samples in ascending order based on event time. Handle censored data appropriately: censoring time affects only the risk set but not the event count.

2. For each time point t_i:

   - Compute d_i, the number of events (e.g., deaths) occurring at time t_i, and n_i, the number of individuals at risk just prior to t_i.
   - Update the survival probability as:

   $$S(t_i) = S(t_{i-1}) \times \left(1 - \frac{d_i}{n_i}\right)$$

3. After iterating through all event times, assign each sample its corresponding survival probability $S(t)$, effectively merging the `efs` (event indicator) and `efs_time` (event time)

variables into a single target variable $S(t)$, which can be used as the target for predictive modeling.

The algorithm handles censoring by only including uncensored cases in the calculation of d_i, but counting censored cases in n_i up to the time of censoring.

3.2 Cox Proportional Hazards Model

The Cox model assumes a hazard function of the form:

$$h(t|X) = h_0(t) \exp(\beta^T X)$$

where $h_0(t)$ is the baseline hazard and β represents feature coefficients. Parameters are estimated via partial likelihood:

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta^T X_i)}{\sum_{j \in R(t_i)} \exp(\beta^T X_j)} \right]^{\delta_i}$$

with $R(t_i)$ being the risk set at t_i, and $\delta_i = 1$ if the event occurred.

3.3 Hybrid Models with Gradient Boosting

We construct five models by combining KM or Cox-based labels with gradient boosting frameworks.

In Kaplan-Meier-type models, the survival probability $S(t)$ is estimated based on the `efs` (event indicator) and `efs_time` (event time). Then, $S(t)$ is used as the target variable, while the feature variables are used as predictors. Three gradient boosting models—XGBoost, CatBoost, and LightGBM—are trained to fit the relationship between features and survival probability. Finally, the trained models are used to predict survival probabilities on the test set.

In Cox-type models, the partial likelihood function of the Cox proportional hazards model is used as the loss function. The models learn and optimize based on the feature variables to find the best model fit. In this category, both XGBoost and CatBoost are employed to perform survival modeling and optimization.

Below is a brief introduction to the gradient boosting methods used:

XGBoost

XGBoost is an efficient implementation of gradient boosted decision trees (GBDT), optimizing:

$$L = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

with regularization $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ to prevent overfitting.

CatBoost

CatBoost is optimized for categorical data. Its innovations include:

- **Target Statistics Encoding:** Encodes categorical features using historical label statistics to avoid leakage.
- **Ordered Boosting:** Ensures unbiased gradient estimation.
- **Symmetric Trees:** Improves prediction consistency and efficiency.

LightGBM

LightGBM enhances speed and memory efficiency via:

- Histogram-based training,
- Gradient-based One-Side Sampling (GOSS),
- Exclusive Feature Bundling (EFB),
- Leaf-wise tree growth.

4. Results and Comparison

4.1 Evaluation Metric: Concordance Index

Due to the specific characteristics of survival analysis datasets, the predictive performance of models is evaluated using the Concordance Index (C-index). The C-index is a commonly used metric in survival analysis that measures the concordance between predicted risks and actual survival times. It is defined as the proportion of all comparable pairs of samples in which the individual with a higher predicted risk experiences the event earlier. The mathematical expression is as follows:

$$\text{C-index} = \frac{\sum_{i,j} I(T_i < T_j \wedge \hat{h}_i > \hat{h}_j \wedge \delta_i = 1)}{\sum_{i,j} I(T_i < T_j \wedge \delta_i = 1)}$$

Where:

- T_i, T_j represent the actual event times (or censoring times).
- $\hat{h}^i, \hat{h}^j$ are the predicted risk scores by the model.
- $\delta_i = 1$ indicates that individual i experienced the event (i.e., was not censored).

A value of 0.5 implies random prediction, while 1.0 indicates perfect accuracy.

4.2 Performance Results

Based on the algorithm descriptions above, five models were trained and fitted to predict survival, with their corresponding C-index results shown in the table below.
From the results, it can be observed that the C-index values of the five models are quite close, ranging from 0.6733 to 0.6757. Among them, the XGB-Cox model performed the best. XGBoost demonstrated stable performance under both the Kaplan-Meier and Cox frameworks, achieving C-index values of 0.6751 and 0.6757, respectively. CatBoost and LightGBM showed slightly more fluctuation in performance, especially under the Cox framework, where they performed somewhat less effectively.

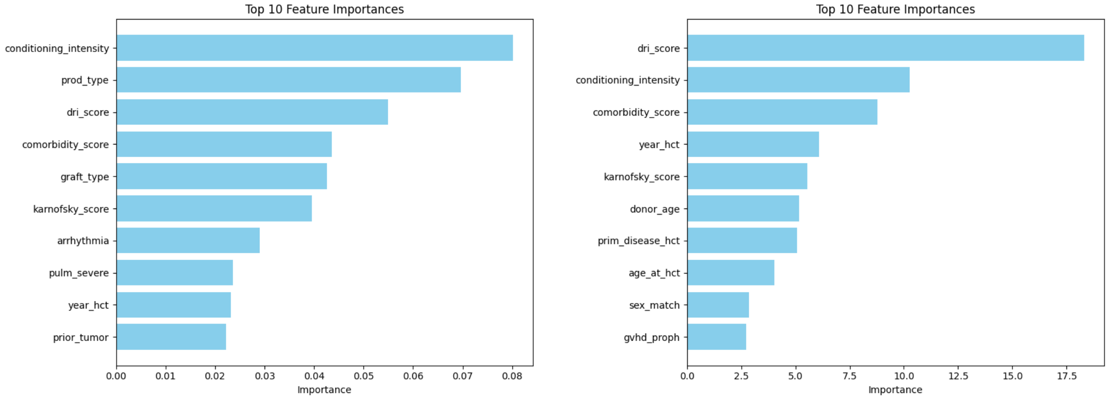
4.3 Feature Importance Analysis

The top important features in different models are shown below. We can get these key findings from feature importance analysis:

- **XGB-Cox:** Top features include `conditioning_intensity`, `dri_score`, and `comorbidity_score`.
- **CatBoost-Kap:** `dri_score` is most important, followed by `conditioning_intensity` and `comorbidity_score`.

Table 1: Model Performance (C-index)

Model	C-index
XGB-Kap Model	0.6751
CatBoost-Kap Model	0.6744
LightGBM-Kap Model	0.6745
XGB-Cox Model	**0.6757**
CatBoost-Cox Model	0.6733



(a) XGB Kaplan Top Importance Features

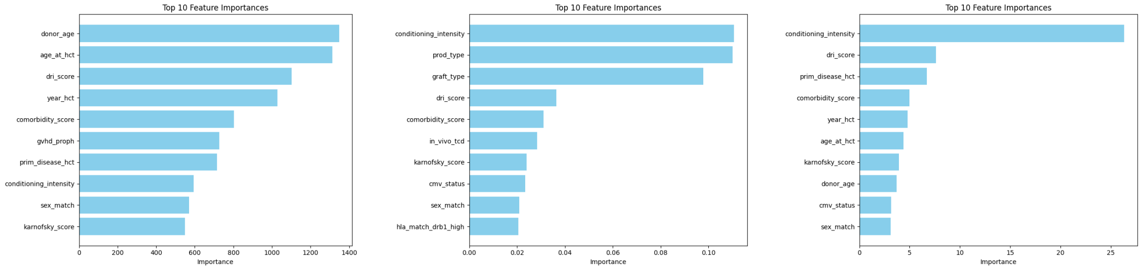
(b) CatBoost Kaplan Top Importance Features

- **LightGBM-Kap:** `donor_age`, `age_at_hct`, and `dri_score` dominate.

Based on the feature importance rankings across the five models, the key factors affecting the survival probability of HCT patients include:

1. **Disease Risk Index (dri_score)**: Ranked highly in all models, indicating that disease risk is one of the most critical factors influencing patient survival. Therefore, patients with high-risk diseases (e.g., AML, ALL) may require more intensive treatment strategies.

2. **Conditioning Intensity (conditioning_intensity)**: Has a significant impact on survival probability, especially ranking high in Kaplan-Meier-based models. Patients with different risk profiles may require different types of conditioning regimens. Myeloablative conditioning (MAC) and reduced-intensity conditioning (RIC) may be suitable for different risk groups.



(c) LightGBM Kaplan Top Importance Features

(d) XGB Cox Top Importance Features

(e) Cat Cox Top Importance Features

Figure 2: Top Importance Features

3. **Comorbidity Score (comorbidity_score)**: Reflects the overall health condition of the patient and is closely associated with survival outcomes.

4. **Year of Transplantation (year_hct)**: May reflect the influence of advances in medical technologies on survival rates.

5. **Performance Status Score (karnofsky_score)**: The physical performance of the patient plays a critical role in determining survival probability.

6. **Age-related Factors (age_at_hct, donor_age)**: The ages of both the patient and the donor are especially important in the LightGBM model, suggesting that age significantly affects transplant outcomes. Younger patients and younger donors may be associated with better prognosis.

5. Conclusion

This study evaluated five models for survival prediction of HCT patients, using Kaplan-Meier and Cox models in combination with XGBoost, CatBoost, and LightGBM. The XGB-Cox model achieved the highest C-index (0.6757), suggesting superior performance in handling censored data and capturing complex feature interactions.

Moreover, consistent patterns in feature importance underscore the critical role of clinical factors such as disease risk, conditioning regimen, comorbidities, and patient/donor age in predicting post-transplant survival. These insights can inform clinical decision-making and personalized treatment planning.For example, particular attention should be paid to the patient's disease risk and comorbidity status before transplantation, while optimizing the conditioning regimen to improve survival rates.