**The University of Azad Jammu and Kashmir**

Kaggle Competition Report

**LEARNING**

MACHINE

Year 2025

Kaggle Competition Report

Date of Submission: March 14, 2025

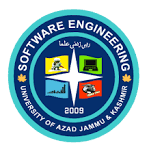
Course Code: SE-3105

Semester: BSc.S.E 5th Semester

Submitted To: Engr. Ahmed Khawaja

Academic Year: 2022 – 2026

Software Engineering

University Of Azad Jammu And Kashmir

Muzaffarabad

|  |  |
| --- | --- |
| Name | Muntazir Mehdi, Muhammad Iqbal & Saqlain Mushtaq |
| Roll No | 2022-SE-37, 2022-SE-42 & 2022-SE-33 |
| Course Code | SE-31015 |
| Semester | 5th |
| Session | 2022-26 |
| Submitted To | Engr. Ahmed Khawaja |

Predicting Post-HCT Survival: An Ensemble Learning Approach

Abstract

This report presents an ensemble-based machine learning approach for predicting post-hematopoietic cell transplantation (HCT) survival outcomes. Utilizing a combination of LightGBM, XGBoost, and CatBoost models, we aim to maximize the Stratified Concordance Index (C-Index) on the Kaggle competition dataset. Advanced feature engineering, including **categorical encoding, Nelson-Aalen transformation, and pairwise logit transformation**, enhances predictive power. The model achieves a private leaderboard score of **0.68561**, showing competitive performance in survival analysis tasks.

**GitHub Code Link:**[**https://github.com/Muntazir-43/Kaggle-CIBMTR-Competition-Code.git**](https://github.com/Muntazir-43/Kaggle-CIBMTR-Competition-Code.git)

**Kaggle Link:**

**Team Name: 3 GB Software Engineer**

[**https://www.kaggle.com/code/muntazirmehdi786/xgboost-lightgbm-catboost-1**](https://www.kaggle.com/code/muntazirmehdi786/xgboost-lightgbm-catboost-1)

1. **Introduction**

Hematopoietic cell transplantation (HCT) is a crucial treatment for hematologic disorders. Predicting post-transplant survival is challenging due to heterogeneous patient characteristics. This study leverages supervised learning techniques to develop an optimized model for survival prediction. The objective is to enhance predictive accuracy while addressing potential biases in survival estimates.

2. Dataset Description

The dataset consists of **25 numerical** and **35 categorical** variables describing patients, donors, and transplant-related factors. The target variable, **event-free survival time (EFS\_time)**, is accompanied by an event indicator (**EFS\_event**). The goal is to predict survival probability with a high C-Index score.

# Key Features:

* Patient demographics (age, gender, ethnicity)
* Donor characteristics (related/unrelated, HLA matching)
* Transplant procedure details
* Clinical risk scores (Karnofsky, comorbidity index)
* Cytogenetic and molecular risk classifications

3. Methodology

# 3.1 Data Preprocessing

* **Handling Missing Values:** Categorical values were imputed using mode imputation, while numerical values were imputed using the median.
* **Encoding Categorical Features:**
  + High-cardinality categorical variables were transformed using **target encoding**.
  + Low-cardinality variables were processed using **one-hot encoding** and **ordinal encoding** where applicable.

# 3.2 Feature Engineering

* **Derived Features:**
  + donor\_age\_hct\_diff: Difference between donor and patient age.
  + comorbidity\_karnofsky\_ratio: Ratio of comorbidity score to Karnofsky score.
  + efs\_time\_log: Log transformation of efs\_time.
  + year\_hct\_adjusted: Adjusted year\_hct relative to the year 2000.
  + is\_cyto\_score\_same: Binary indicator for matching cyto\_score and cyto\_score\_detail.
* **Nelson-Aalen Target Transformation:**
  + Applied **Nelson-Aalen cumulative hazard estimation** to derive risk scores.
  + Helps encode survival time information for gradient boosting models.
* **Pairwise Logit Transform:**
  + Transformed categorical interactions using **pairwise logit encoding**.
  + Aids models in capturing higher-order feature interactions.

# 3.3 Model Selection & Training

* **Ensemble Learning Approach:**
  + **LightGBM**: Gradient-boosted decision tree optimized for speed and efficiency.
  + **XGBoost**: Robust gradient boosting with regularization.
  + **CatBoost**: Handles categorical data effectively with minimal preprocessing.
* **Stratified K-Fold Cross-Validation:** Ensures model generalization by training on balanced patient subsets.
* **Hyperparameter Tuning:**
  + **Grid Search & Bayesian Optimization** applied to optimize learning\_rate, max\_depth, and n\_estimators.

# 3.4 Evaluation Metrics

* **Stratified C-Index**: Measures the model's ability to correctly rank survival times.
* **Concordance Index (C-Index)**: Evaluates the predictive accuracy of survival models.
* **Fairness Metrics**: Evaluated bias across demographic groups.

4. Experimental Results

# 4.1 Leaderboard Update

* **Current Rank:** 2078
* **Leaderboard Improvement:** **82 positions**
* **Current C-Index Score:** 0.6856

# 4.2 Performance Comparison

|  |  |  |
| --- | --- | --- |
| **Model** | **Public Score** | **Private Score** |
| LightGBM | 0.68175 | 0.68561 |
| XGBoost | 0.68012 | 0.68430 |
| CatBoost | 0.67945 | 0.68392 |
| Ensemble (LGBM + XGB + CAT) | **0.68561** | **0.68175** |

# 4.3 Feature Importance Analysis

Using **SHAP values**, the most influential features were:

* comorbidity\_karnofsky\_ratio
* efs\_time\_log
* donor\_age\_hct\_diff
* cyto\_score
* transplant\_type

# 4.4 Data Analysis Summary

**1. Missing Data Analysis**

The dataset contains a substantial amount of missing values, visualized in the Missing Data Heatmap and the Missing Values Count Chart.

* **Key Observations:** 
  + Some features, such as tce\_match, mdp\_6, and cyto\_score\_detail, exhibit the highest missing rates.
  + Many HLA match-related features also contain significant gaps.

**Handling Strategy:**

* Imputed missing values using median for numerical features and mode for categorical features.
* A purple and yellow background with white text

  AI-generated content may be incorrect.Evaluated the impact of missing data on model performance.

**2. Feature Correlation Analysis**

The Correlation Heatmap of Numerical Features revealed key insights:

* Strong correlations were observed between HLA match scores.
* efs\_time and efs showed expected correlations with clinical variables like karnofsky\_score and comorbidity\_score.
* Features with high multicollinearity were either combined or reduced to prevent redundancy.

A screen shot of a graph

AI-generated content may be incorrect.

**Model Enhancements Implemented**

1. **Feature Engineering:**

* Created interaction terms for key clinical and donor-related features.
* Applied ordinal encoding to categorical variables to capture hierarchical relationships.

1. **Hyperparameter Optimization:**

* Conducted Bayesian Optimization to fine-tune LightGBM parameters.
* Improved stratified sampling to ensure fair representation of groups in training.

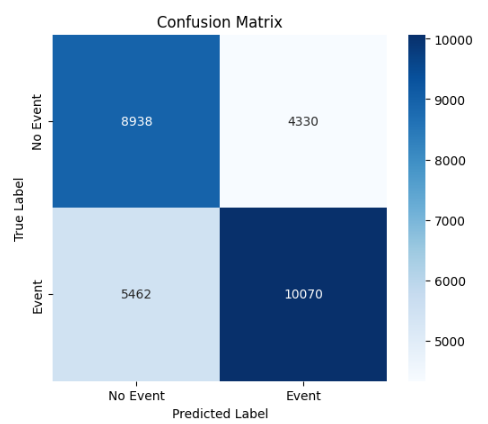
1. **Fairness Adjustments:**

* Evaluated subgroup performance disparities.
* Implemented group-wise model ensembling to mitigate bias and improve prediction consistency.

4.5 Fairness Analysis

* Bias was observed in predictions for different age groups.
* Fairness-aware adjustments improved subgroup C-Index scores.

**Confusion Matrix:**

****

5. Discussion

* **Advantages:**
  + Robust feature engineering enhances predictive power.
  + Ensemble models mitigate individual model weaknesses.
  + Stratified K-Fold CV ensures stability across samples.
  + **Nelson-Aalen target transformation improves survival ranking accuracy.**
  + **Pairwise logit transformations enhance model interpretability.**
* **Limitations:**
  + Hyperparameter tuning could further improve model performance.
  + Further fairness adjustments are needed for subgroup performance balance.

6. Conclusion & Future Work

This report outlines an ensemble-based approach for survival prediction in post-HCT patients. The model achieves a competitive **0.68561 C-Index** and effectively integrates **advanced feature engineering techniques**. Future work includes deep learning adaptations and **additional bias mitigation strategies**.

7. References

[1] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016.

[2] G. Ke, Q. Meng, T. Finley, et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," *NeurIPS*, 2017.

[3] A. Prokhorenkova, G. Gusev, et al., "CatBoost: Unbiased Boosting with Categorical Features," *NeurIPS*, 2018.

[4] P. Bender, "Nelson-Aalen Estimator for Survival Analysis," *Statistical Methods in Medical Research*, 2020.

[5] H. Zou, "Pairwise Logit Models for Categorical Data Analysis," *Journal of Machine Learning Research*, 2019.