**Prediction of Coronary Artery Disease mortality with MIMIC III**

**BIS568 Final Project Report**

**Group members: Zhiyuan Cao, Haoting Chen, Ziqing Ji**

## 1. Introduction

### Problem Specification

Coronary artery disease (CAD) is the most common heart condition. It is caused by cholesterol deposits in the major blood vessels that supply the heart, coronary arteries. Over time after a CAD diagnosis, inflammation will cause the deposits to grow, leading to hardened and narrowed blood vessels that are incapable of sending enough blood, oxygen, and nutrients to the heart. Eventually, the complete blockage of the coronary arteries will lead to a heart attack.

CAD is the leading cause of death in the United States, accountable for 1 in 4 deaths in the US (approximately 610,000 deaths every year). It is also the third leading cause of death in the entire world, accountable for approximately 17.8 million deaths globally every year.

Despite the fact that there is no cure for CAD, appropriate and adequate treatment can help symptom management and reduce the chances of serious events such as heat attacks. In particular, patient subgroups with high short-term mortality rates deserve extra attention and tailored treatments from caregivers (such as close family members and nursing homes) and healthcare providers (such as hospitals and specialty clinics). In order to reduce the predictable and preventable deaths, this project intends

to provide a machine learning solution for the 3-year mortality prediction of patients diagnosed with CAD in intensive care.

### Data Preparation

Using MIMIC III database, a subset of the condition table with records of coronary artery disease was loaded. After that, a subset of the person and death table where person_id is present in the previously obtained condition table was loaded. A subset of the visit and measurement table where visit_id is present in the previously obtained condition table was loaded. Concept table was loaded based on concept_id to annotate the measurement table. Then, the loaded tables were merged to obtain 840 visit-level entries. Finally, the outcome variable “death” was created (death = 1 if death_date – visit_date ≤ 3 years, the patient died in 3 years after an intensive care diagnosis of CAD; death = 0 if death_date – visit_date ≤ 3 years or death_date = NA, the patient died after 3 years/didn’t die).

## 2. Methods

### Model Selection

Two models were implemented in this project: XGBoost and Random Forest, and their performance was compared. XGBoost allows the prediction of a single target variable, which is “the mortality rate or status” in our dataset. It runs relatively fast and is able to predict with a relatively high accuracy. On the other hand, Random Forest consists of many decision trees, which renders an accuracy that is much higher than using each of the individual trees. The accuracy of these two models ended up to be about 79% for XGBoost, and about 74% for Random Forest.

### Feature Selection and Difficulties

Feature selection was accomplished mainly through researching the significant factors of CAD and trying different combinations of features. In the end, we kept the combination that generated the highest accuracy scores for both models. One of the main difficulties in feature selection was that there was a lot of missing data for some of the variables, especially measurements. Therefore, when trying different combinations of features, it was important to test how much remaining data there is after removing rows with missing data.

## 3. Results

### ROC-AUC curves, Pr-AUC curves and Calibration curves

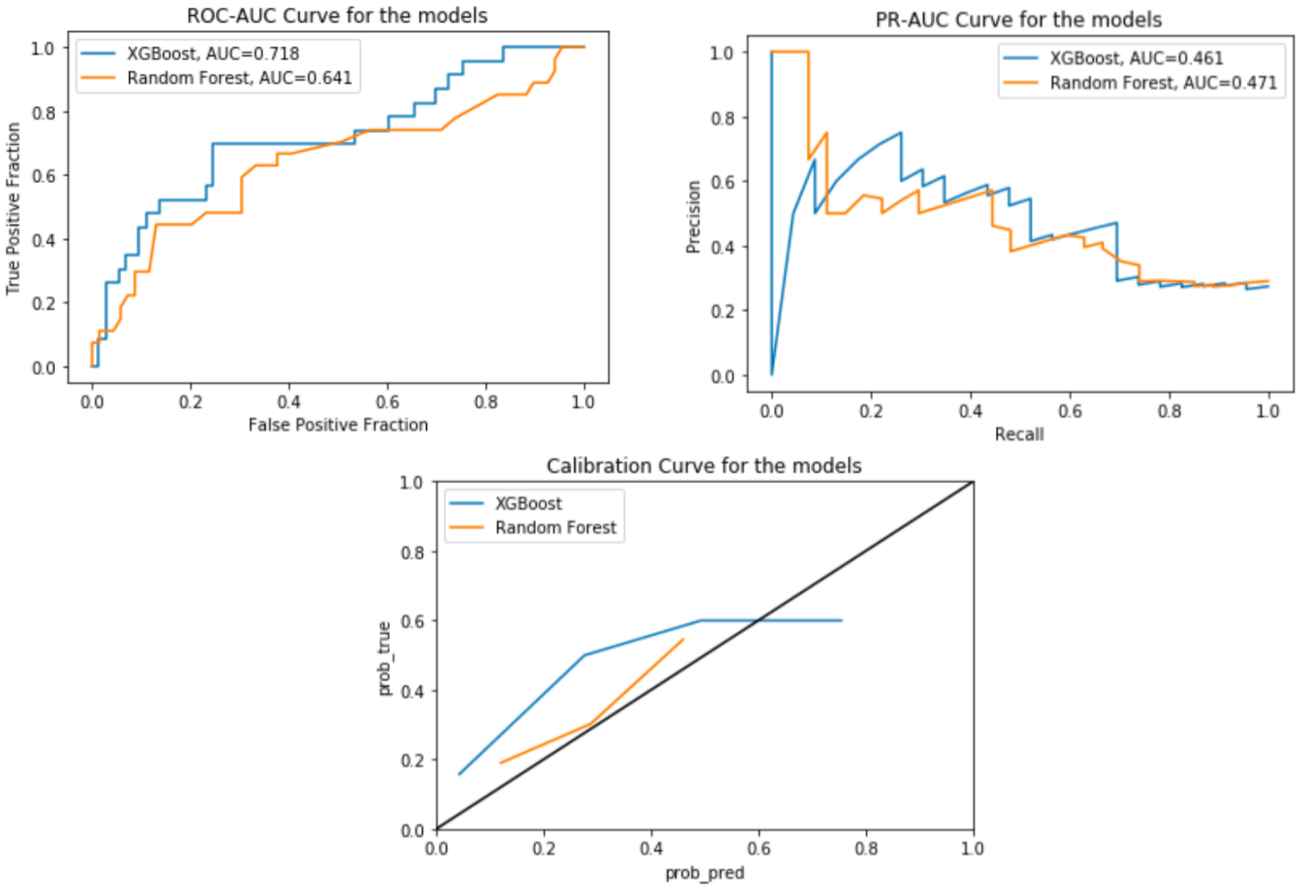


**Fig. 1.** Performance of the models

Fig.1 shows the ROC-AUC, PR-AUC and calibration curves for our models. It looks like the XGBoost model is performing better than the Random Forest model in terms of AUC for predicting Coronary Artery Disease mortality. In this case, the XGBoost model has an AUC of 0.718 while the Random Forest model has an AUC of 0.641, which is lower than the XGBoost model's AUC and suggests that it is not performing as well.

It is important to note that ROC-AUC is just one metric and it is important to consider other evaluation metrics as well. So we plot the PR-AUC curves for our models.

To evaluate the accuracy of a model's probability predictions, we plot the calibration curve to visualize the degree to which the predicted probabilities align with the true outcomes. We can see that both model are close to the diagonal line that represents perfect calibration

**Confusion Matrix**

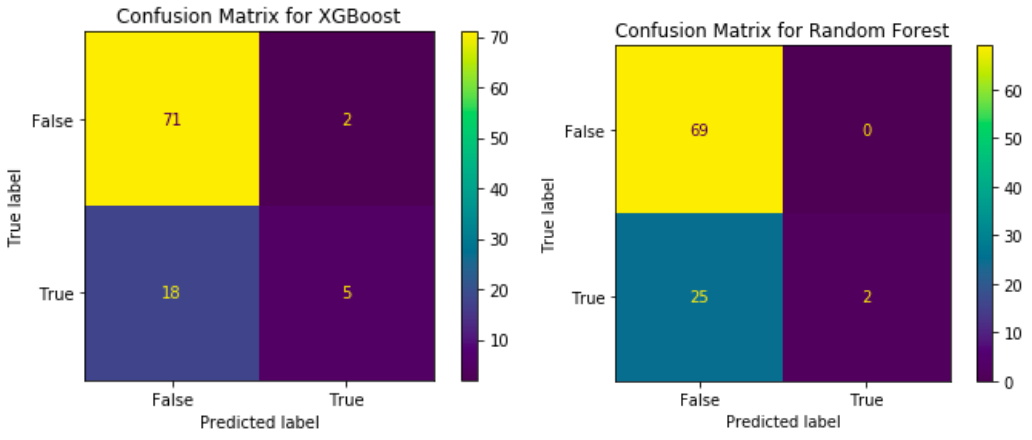


**Fig. 2.** Confusion Matrix of the models

The confusion matrices on the test set for our models are shown in Fig. 2. Generally the performance is satisfactory. The False negative rate is a little bit higher in our models.

## Interpretability and Explainability

Fig. 3 is the SHAP plot for our model. From these figures, we can see that Plateles in blood, Prothrobin time and BMI are the three most important features in our dataset. In contrast, race and gender are factors that are not that important.

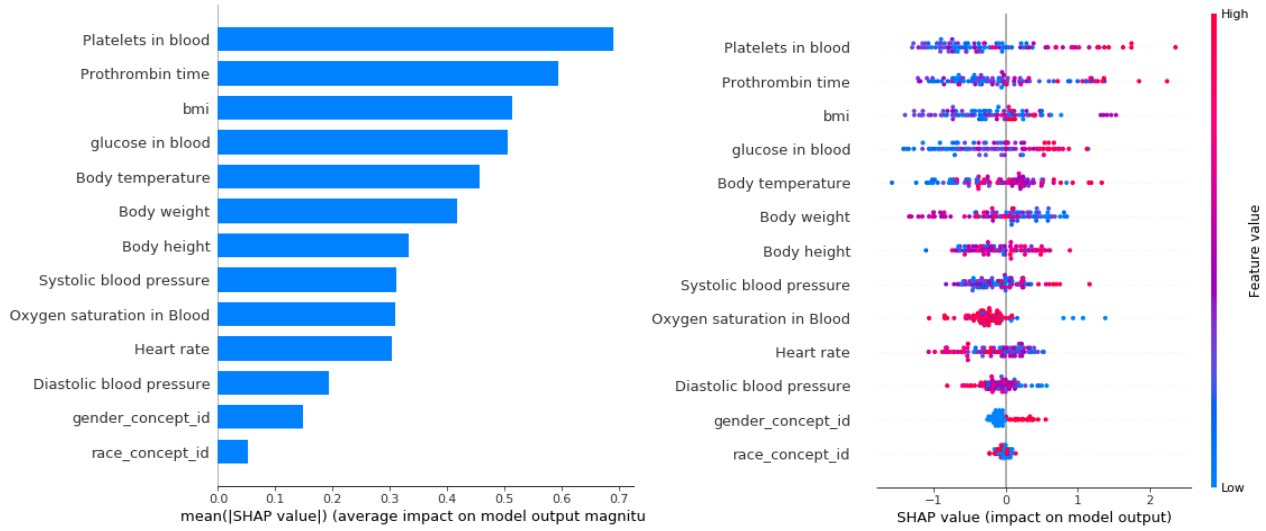


**Fig. 3.** SHAP summary plots of the models

## Bias Analysis

We split the dataset by races: white and non-white. From Fig. 4, it turns out that the slight differences on performance comes from the unbalanced dataset: Both algorithms perform a little bit worse on minority classes than majority classes.

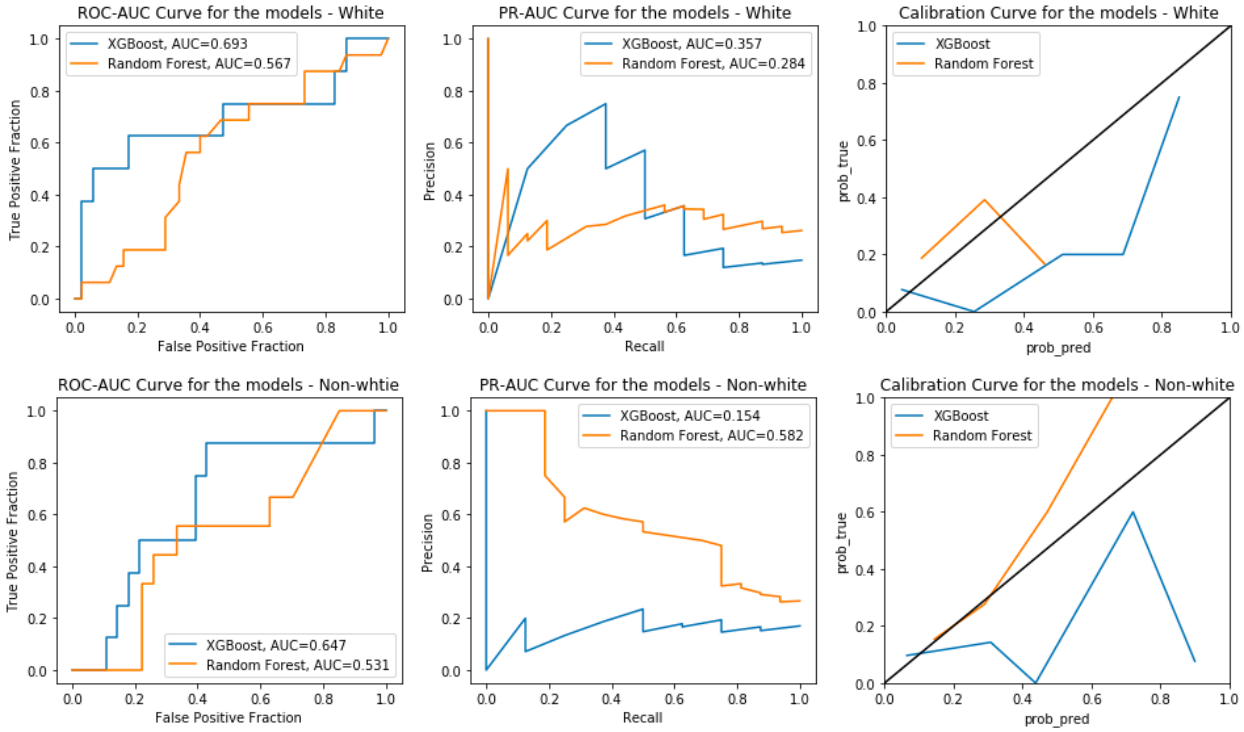


**Fig. 4.** Bias analysis for races - White and Non-white

## 4. Discussion

### Model Results

The AUC of our ROC curve is relatively high and suggests that it is able to distinguish between positive and negative examples relatively well (i.e., patients who did or did not die from Coronary Artery Disease).

On the other hand, the False negative rate is a little bit higher in our models. The PR-AUC curves also have lower AUC values. Both phenomena are caused by imbalanced classes: The positive class is rarer than the negative class in the dataset.

### Evaluation, Implementation, and Dissemination Plans

Regarding our evaluation plan, firstly, it is important to continue to increase the accuracy scores of the models, through implementing different algorithms or models. Secondly, testing with new testing data could show how generalized the model is and how applicable it is with other real-world data. More up-to-date data could be collected from local hospitals or specialized clinics. Thirdly, it is also necessary to ask for patient consent to use their data in testing our models. For example, it is noticed that age data was hidden to protect privacy in the MIMIC dataset. Age is a significant factor in development of CAD, therefore, adding it to our model might make a difference in the accuracy scores. Therefore, asking for consent to use specific kinds of data is also an important step in evaluating and testing our model.

Regarding implementation plan and dissemination strategies, cooperation with local hospitals and specialized clinics is necessary. After accuracy is improved to a certain level, it could be available online for CAD healthcare providers to use. The model could be applied as a CDS tool, supporting decision making in diagnosis, treatment planning, and so on, which could help prevent predictable mortality.

## 5. Conclusion

In conclusion, our analysis using the MIMIC III dataset found that the XGBoost model was more effective for predicting mortality due to coronary artery disease compared to the random forest model. The XGBoost model had an accuracy of 79%, while the random forest model had an accuracy of 74%. While these results are promising, it is important to note that there is still room for improvement and further research is needed to fully understand the capabilities and limitations of these

approaches. Overall, our study suggests that machine learning models can be useful tools for predicting mortality due to coronary artery disease.