LIFEHACK DATATHON 2020

Presentation By Team Exception:

Choo Su Lyn

Zhou Kai Jing

INTRODUCTION

The COVID-19 pandemic has resulted in widespread implications on the healthcare industry as healthcare systems are overloaded with the number of patients. As such, it is pertinent for healthcare systems prioritise patients that are more likely to succumb to the virus so as to reduce the mortality rate of COVID-19. Thus, machine learning methods were explored to predict and potentially identify such patients so that early intervention can be done to reduce their likelihood of death.

PRESENTATION OUTLINE

DATA PREPARATION

Data Wrangling, Data Cleaning & Feature Selection

MODEL SELECTION

Identifying Potential Models, Model Testing & Establishing Baselines

HYPERPARAMETER TUNING

Model Improvement & Optimization

MODEL INTERPRETATION

Model Assumptions & Evaluation Metric

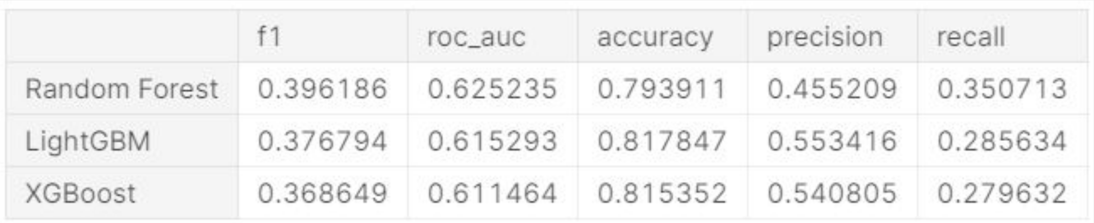
DATA PREPARATION

Our first steps in data preparation involves identifying and creating the target feature. The 'is_dead' feature created was based on 'death_date' of patients and is used to represent the state of a patient's mortality.

Hence, the goal of this project is targeted towards the prediction of patients' mortality based on other features such as age, gender, underlying background diseases, etc.

In the process of data wrangling, we have also identified that the the features included on different symptoms of a patient have a lot of missing data values. In addition, since they are one hot-encoded values, it is difficult to perform data imputation and might potentially skew our final results. Hence, the symptoms features were excluded from our final dataset.

MODEL SELECTION



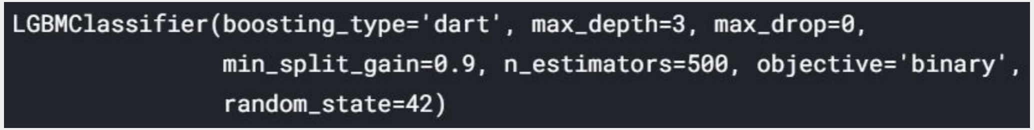
	f1	roc_auc	accuracy	precision	recall
Random Forest	0.396186	0.625235	0.793911	0.455209	0.350713
LightGBM	0.376794	0.615293	0.817847	0.553416	0.285634
XGBoost	0.368649	0.611464	0.815352	0.540805	0.279632

To start off with our model building, three extremely popular and industrial-grade models were selected to establish their baseline performances for comparison.

Random Forest and Extreme Gradient Boosting methods have been popular tools in the Machine Learning toolbox due to their ability to perform well across many scenarios. The Light Gradient Boosting method has also seen a recent surge in popularity due to its efficiency, making it a practical choice given the large dataset that we are working with.

Comparing across the different baseline metrics, Light Gradient Boosting was our choice of model due to its high accuracy, its efficiency in terms of compute time, as well as its relatively high ROC AUC score (which indicates its ability to accurately classify across different thresholds set in the model).

HYPERPARAMETER TUNING



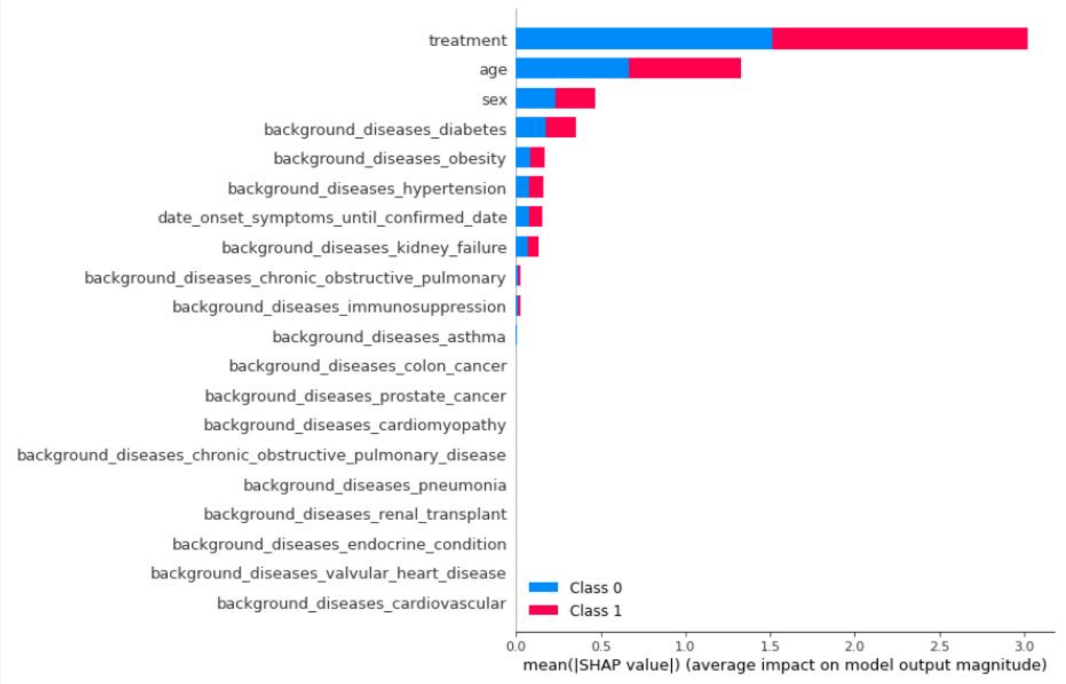
```
LGBMClassifier(boosting_type='dart', max_depth=3, max_drop=0,
               min_split_gain=0.9, n_estimators=500, objective='binary',
               random_state=42)
```

In order to improve our Light Gradient Boosting (LightGBM) model, we underwent a series of hyperparameter tuning before arriving at the best version of our model.

First, 'boosting_type' and 'learning_rate' are two parameters that allows us to improve the overall accuracy of the model by determining the best type of tree booster and the best learning rate to adopt. A lower learning rate, coupled with increased tree estimators for learning allows for better accuracy overall whereas DART booster allows tree dropouts along the training phase to combat overfitting by tuning the dropout rate in the subsequent step.

Finally, the tuning of 'max_depth' and 'min_split_gain' also aims to combat model overfitting by limiting both the tree complexity and the level of tree growth. This results in our final model with the hyperparameters shown above.

MODEL INTERPRETATION

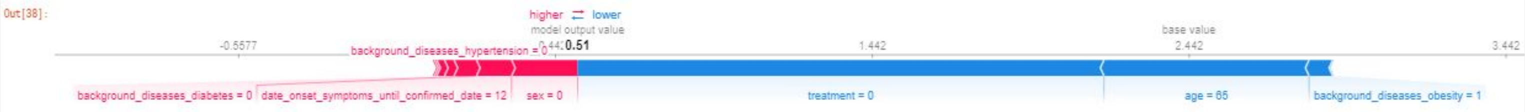


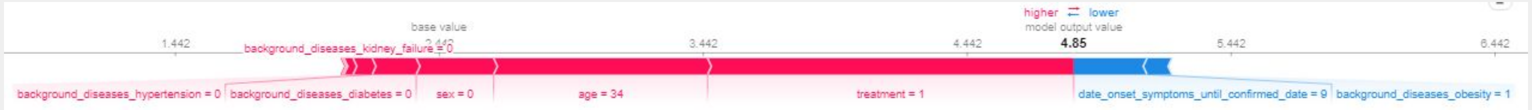
Using SHAP, we can identify the most relevant features forum our LightGBM model based on the ranked feature importance shown on the left.

From our model, the most important features in the model are determined to be treatment, age, gender, and whether a patient has pre-existing diseases such as diabetes, obesity, hypertension and kidney failure.

MODEL INTERPRETATION

Out[38]:





For a quick understanding of the chart above, features identified in Red represent features that contribute towards increasing the predicted mortality value of a particular patient, whereas features highlighted in Blue contributes towards a reduction in the reduced mortality of the patient.

The further visualization on two of our patient samples using SHAP shows us that the Male patients (indicated by sex = 0) could face a higher mortality rate when infected with COVID-19. In addition, pre-existing conditions such as diabetes, hypertension and kidney failure also tend to contribute towards increasing the probability of mortality in a patient.

These additional knowledge adds to our current knowledge that patients with very low or high age, as well as those not receiving treatment, are generally more susceptible to death when infected. These, however, are not accurately reflected in the visualizations of the small samples above.

INSIGHTS & REFLECTIONS

Despite all our model improvements, more can be done to obtain a better understanding of the relationship between the target feature and the other features. One way to achieve this is the exploration of methods such as Synthetic Minority Oversampling Technique (SMOTE) to combat the class imbalance problem in the target feature.

In terms of interpretation, we should also always keep in mind that while SHAP gives us the ranked feature importance, it does not actually indicate any causality between the target feature and the ranked features.

As the COVID-19 pandemic unfolds itself over time, more and more patient data will become available. This will allow us to better understand the disease and reach closer to the ground truth on what actually contributes to or detracts from the conditions of infected patients.

THANK YOU

Analysis By Team Exception:
Choo Su Lyn
Zhou Kai Jing