**Boosted Tree**

We conducted gradient boosted tree model to build prediction model. The model introduction, modeling process, and results interpretation are shown as below.

**Model Introduction**

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient boosted tree fits a sequence of trees so that each tree concentrates on misclassified records from previous tree and gets improvements each time. In other words, gradient boosted tree is built in a stage-wise fashion and it can generalize them by allowing optimization of loss function.

Gradient boosting tree is especially useful when the labels of records are imbalanced. For example, for death prediction, most labels are non-death, but we are interested in the labels of death. Boosted trees can increase the accuracy of rare class of special interest. Basic classifiers are tempted to classify cases as belonging to the domain class. And misclassifications with the single best-pruned tree can be those rare class of special interest, which is what boosting algorithm concentrates on. Although we will rebalance the data first to make two classed balanced and don’t have severe imbalanced classes problem, it is still good to try gradient boosted tree model and see whether it can improve the prediction performance.

**Modeling Process**

First of all, we split the original data into two parts: train data and test data. Train data was for training the model and test data was to evaluate the model to see its prediction performance about unseen data. We used threshold of 75/25 for splitting train and test data.

We found that we come across imbalanced class distribution. The number of observations belonging to death is significantly lower than those belonging to non-death – we saw that the death rate of the original data set was about 5.7%. In this situation, the predictive model developed using conventional machine learning algorithms could be biased and inaccurate.

There mainly to ways to deal with imbalanced data, under-sampling and oversampling. Under-sampling balances the dataset by reducing the size of the abundant class. Oversampling is used when the quantity of data that we care about (e.g. death cases) is insufficient. It tries to balance dataset by increasing the size of rare samples. Rather than getting rid of abundant samples, new rare samples are generated by using bootstrapping. Considering the need to reduce information loss, we decided to use bootstrapping approach to rebalance the train data.

For the model tuning, we used Grid Search function to help us find the optimized parameters of gradient boosted tree model, and finally used n\_estimators=75, max\_depth=4, learning\_rate=0.5.

Finally, we used test data (92,659 observations) to evaluate the model performance, including confusion matrix, F1 score, and top 10 important features. The results were shown and interpreted in the following part.

**Results Interpretation**

1. **Confusion matrix**

According to the confusion matrix of gradient boosted tree model, we concluded that the overall accuracy (was 94%.

The precision (was 51%, which meant that for the cases we predicted to be dead, 51% of them were true death cases.

The recall (was 93%, which meant that for the cases that were dead, our model can predict them correctly with accuracy of 93%.

Since it was bad to make mistakes, we further analyzed the mistakes and costs of the gradient boosted tree model. There are two different kinds of mistakes: false positive and false negative. False positive means that we regard a health person as an at-risk case. This can lead to unnecessary procedure such as quarantine and bring stress to people who are predicted as high risk but actually not. Another problem is false negative, which can be more serious because it means that the model can’t predict the high-risk cases and regard them as safe. In such condition, people with high risk don’t get treatment, infect others and eventually die.

According to the confusion matrix, we concluded that the false positive (was 49%, and the false negative ( was 7%. Gradient boosting tree model may make a bigger mistake on predicting a health person as a high-risk one, but it was not a big deal because unnecessary quarantine didn’t cost too much. Fortunately, our false negative was very low which meant that we can predict at-risk population with slight mistakes. This is much more important because we can not only save lives by providing early treatments on those high-risk population, but also ensure the safety of healthy population by getting rid of being infected.

**Table 1 Confusion matrix of the boosted tree**

|  |  |  |
| --- | --- | --- |
|  | **Predicted non-death** | **Predicted death** |
| True non-death | 82624 | 4766 |
| True death | 374 | 4895 |

1. **F1-score**

Furthermore, we analyzed the F1 score of each class, especially class 1 which represented death. The results were shown as below.

F1 score (2 × is the harmonic mean of Precision and Recall, and it gives a better measure of the incorrectly classified cases than the accuracy metric. We used the harmonic mean since it penalized the extreme values. Accuracy can be used when the class distribution is similar while F1-score is a better metric when there are imbalanced classes. Thus, in this case, it will be more reasonable to look at F1-score.

According to the results, we concluded that the F1-score of class 1 (death) was 0.66, which was slightly lower than that of our benchmark model - Naive Bayes.

**Table 2 F1 scores of the boosted tree**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **precision** | **recall** | **f1-score** | **support** |
| 0 | 1.00 | 0.95 | 0.97 | 87390 |
| 1 | 0.51 | 0.93 | 0.66 | 5269 |
| accuracy |  |  | 0.94 | 92659 |
| macro avg | 0.75 | 0.94 | 0.81 | 92659 |
| weighted avg | 0.97 | 0.94 | 0.95 | 92659 |

1. **Feature importance**

Finally, we cared about the important features of model that influenced the most. The importance of feature measured how important a feature was in the context of a model. That is, it measured how much (based on the training data) the accuracy of a model would decrease if that feature were removed.

手机截图图社交软件的信息

描述已自动生成

**Figure 1 Top 10 most important features**

According to the plot below, we can see that the top10 most important features are 'CHEST\_Status\_A', 'CHEST\_Other\_Nos\_B', 'CHEST\_Airway\_Lungs\_A', 'GENRL\_UNSP\_Service', 'CVASC\_Heart\_Rhythm\_A', 'ENDOC\_MET\_Metabolic\_A', 'Age\_75\_99', 'ENDOC\_MET\_Diabetes', 'CVASC\_Other\_Nos\_B', and 'CVASC\_Cardiac\_B'.

We made a simple conclusion on these important features: a person aging from 75 to 99 with some diseases such as cardiac and heart problems, lungs problems, and diabetes may have a higher risk to die from COVID-19. We will further discuss this and provide deeper insights in the Section *Insights and Suggestion*.

**Insights and Suggestion**

Comparing all predictive models and top10 important features of each model, we concluded top10 common important features that had the highest predictive influence on whether a person was at a high risk of death from COVID-19. The features and their descriptions were listed as below.

**Table 3 Top10 common features of importance**

|  |  |
| --- | --- |
| **DGL\_3\_Extend** | **Description** |
| CVASC\_Cardiac\_B | Cardiac diseases such as heart failure |
| CVASC\_Other\_Nos\_b | Mixed diseases include lung, heart, blood pressure, blood vessel, etc. |
| ENDOC\_MET\_Diabetes | Diabetes and concomitant disease like diabetic chronic kidney disease |
| ENDOC\_MET\_Metabolic\_A | Metabolic abnormality such as obesity |
| GENRL\_UNSP\_Service | Aftercare, counseling, and medical exam |
| Age\_75\_99 | Elder population |
| CVASC\_Heart\_Rhythm\_A | Irregular heartbeats such as tachycardia |
| CHEST\_Airway\_Lungs\_A | Lung and respiratory abnormalities |
| GENRL\_UNSP\_Other\_Nos\_A | Mixed diseases include infection of liver, immunodeficiency, etc. |
| CHEST\_Status\_A | Lung and respiratory abnormalities |

We drew a patient persona that have a higher risk of dying from COVID-19:

S/he is an elder person aging from 75 to 99 or people of any age with existing health problems. Those health problems are cardiac and heart rhythm diseases, diabetes and the chronic disease brought by it such as diabetic chronic kidney disease, metabolic abnormality such as obesity, lung and respiratory abnormalities, blood pressure and vessel problems, liver failure, and diseases of the immune system. S/he may also rely on aftercare and medical exam.

Our model results were completely consistent with those of CDC and other scientific sources and can be cross verified with each other.

According to the report of centers for disease control and prevention (CDC)[[1]](#footnote-1), older adults and people of any age who have serious underlying medical conditions might be at higher risk for severe illness from COVID-19.

To be specific, those at high-risk for severe illness from COVID-19 are:

* People 65 years and older
* People who live in a nursing home or long-term care facility

People of all ages with underlying medical conditions, particularly if not well controlled, including:

* People with chronic lung disease or moderate to severe asthma
* People who have serious heart conditions
* People who are immunocompromised, for example cancer treatment, smoking, bone marrow or organ transplantation, immune deficiencies, poorly controlled HIV or AIDS, and prolonged use of corticosteroids and other immune weakening medications
* People with severe obesity (body mass index [BMI] of 40 or higher)
* People with diabetes
* People with chronic kidney disease undergoing dialysis
* People with liver disease

Apart from the key features above, our other important features were also verified, such as hypertension (high blood pressure), which was also thought as an underlying medical condition that drove people to be at risk for sever illness from COVID-19 according to government of Canada[[2]](#footnote-2).

Harvard Medical School claimed that about 10% of patients with pre-existing cardiovascular disease (CVD) who contract COVID-19 will die, compared with only 1% of patients who are otherwise healthy. Increased risk has also been seen in people with high blood pressure (hypertension) and coronary artery disease (CAD)[[3]](#footnote-3). It also proved that cardic disease was the most important feature in almost all predictive models we generated.

Figuring out people who are at higher risk for severe illness from COVID-19, we provided three suggestions to protect these people.

First of all, we should make sure people with high-risk stay at home even when we are making re-opening efforts. We should be careful and cautious to allow those people to go back to workplace.

Moreover, government may consider employing them and pick up their salary until a vaccine is made available to them. According to the research, personal financial condition also affected the probability of disease[[4]](#footnote-4). It is important to ensure their financial power to enhance their ability to resist risks.

Finally, if a vaccine is available, we should prioritize the at-risk population for early reception of the vaccine. Then we can consider to making it available for the population with low risk.

1. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html [↑](#footnote-ref-1)
2. https://www.canada.ca/en/public-health/services/publications/diseases-conditions/people-high-risk-for-severe-illness-covid-19.html [↑](#footnote-ref-2)
3. https://www.health.harvard.edu/blog/how-does-cardiovascular-disease-increase-the-risk-of-severe-illness-and-death-from-covid-19-2020040219401 [↑](#footnote-ref-3)
4. https://www.kff.org/disparities-policy/issue-brief/low-income-and-communities-of-color-at-higher-risk-of-serious-illness-if-infected-with-coronavirus/ [↑](#footnote-ref-4)