

Thank you Lucy!!

Now that we have decided on our final model we thought about how to optimize it.

Having irrelevant input features can mislead the model resulting in worse predictive performance.

So, we performed univariate analysis using chi-squared test to study the relationship between the underlying condition groups and death from covid.

The result shows that all the underlying conditions are significantly associated with death.

(Next Slide)

Next, we performed multivariate analysis using logistic regression. The result shows similar trends to those of the univariate analysis, except for pregnancy and cardiovascular.

Also, comparing the weights of the features we see that Pneumonia, age, gender and renal chronic are the most features associated with death.

(Next Slide)

After that we used feature selection algorithm called, Recursive feature elimination.

This algorithm selects a subset of the most relevant features.

It starts with all the features in the training model and removes the insignificant features until the desired result.

You can see in the graph to the left how the performance of the model changed with the number of input features.

The accuracy of the model is almost constant after seven features.

The algorithm selected the most relevant features which are .....

So we compared our final model with these 7 features and with all the features.

As you can see the accuracy is slightly better in the full model so we decided to keep all the features to not lose any information.

(Next Slide)

Next we used shap library in python for the model interpretation.

The red graph shows how the model classified a patient who has pneumonia, and his is in the senior group.

This patient was classified as high risk patient and we can see that having pneumonia and his age played the greater factors in the prediction of his class.

The other patient is also senior, and has obesity. He was classified as low risk because he doesn't suffer from pneumonia.

So pneumonia is the most important feature and it explains more than 30% of the model.

(Linda's recording)