PREDICTING EARLY ICU ADMISSION BASED ON COVID-19 DIAGNOSIS OF PATIENTS IN BRAZIL.

Sheetal Shajan, Angelene Leow

BACKGROUND MOTIVATION

As the world continues to struggle to contain the ongoing novel virus that is COVID-19, the major obstacle faced by healthcare institutions is the unpredictability in demand for healthcare resources such as ventilators and ICU resources to treat critically ill patients. This obstacle can be mitigated if a machine learning model is able to predict correctly if a patient with covid-19 will be admitted to the ICU hours before it happens. For our project, we have chosen to look at the Covid-19 cases specifically from Hospitals Sírio-Libanês, São Paulo and Brasilia in Brazil.

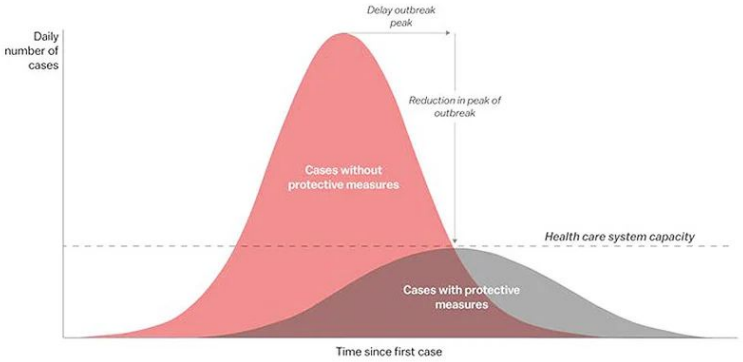

Figure 1: COVID-19 Sirio-Libanese Dataset

THE DATASET:

The features in the dataset included patient demographic information, disease history, blood results, various measurements of vital signs and whether or not they were admitted to the ICU (**0** for no admission, **1** for admission).

MODELLING:

Our goal is to predict with accuracy and transparency. The following steps were employed for modelling and prediction:

1. ExploratIon
2. Data preprocessing
3. Modelling
4. Visualizing results

1. We randomly split our dataset into training and testing with ratios of 75/25. No outliers were present in the dataset because the numeric variables in the data were already scaled.
2. We then filled in the median value to replace missing data. We preprocessed the categorical variables by creating binary variables for each unique categorical value. This was done using Sklearn's One Hot Encoder.
3. We ran our training and testing set through a DummyClassifier to explore the distribution of data.

```
In [20]: dc = DummyClassifier(strategy='prior')
         dc.fit(X_train,y_train)

Out[20]: DummyClassifier(strategy='prior')

In [160]: dc.score(X_train,y_train)

Out[160]: 0.7297921478060047

In [159]: dc.score(X_test,y_test)

Out[159]: 0.7564766839378239

In [22]: dc.predict(X_train)

Out[22]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

Figure 2. Dummy classifier score showing patients admitted to ICU (Value = 1) and not admitted to ICU (Value = 0) in our training vs test set.

Figure 2 shows an uneven distribution for both the training and testing set. Dummy classifier predicts the most frequent value in the target variable for all samples. From this, we can see that there is an approximately 75:25 ratio of patients not admitted to ICU: patients admitted.

A random forest model was run with five-fold cross-validation on the training set and it gave a validation accuracy of 86.01%. Cross-validation calls fit and predict on our model five times (in this case) by randomly splitting the data with test and validation. The model gave a test accuracy of 89.12%.

4. To dive deeper into our analysis, we looked into the sensitivity (which is a measure of how many of the actual positive examples we correctly identified) and precision of our model.

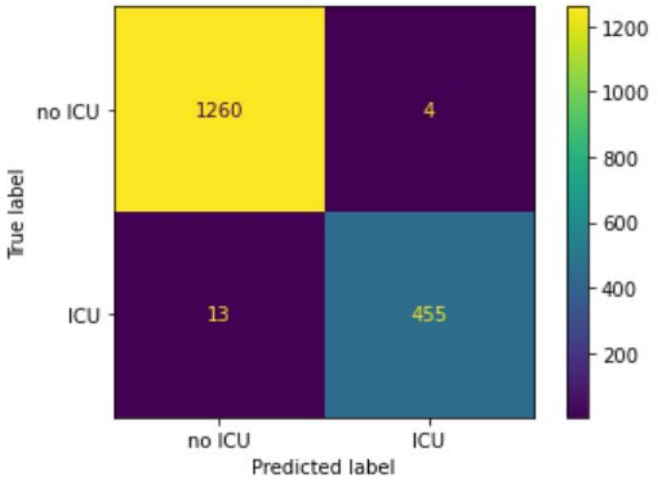

Figure 4. Error matrix using the random forest model on training data.

Figure 4 shows our model's prediction counts on the training data. Out of all 468 patients who were admitted to the ICU, our model predicted 97% of the patients correctly. For the patients not admitted to the ICU, our model shows a 99.8% accurate prediction.

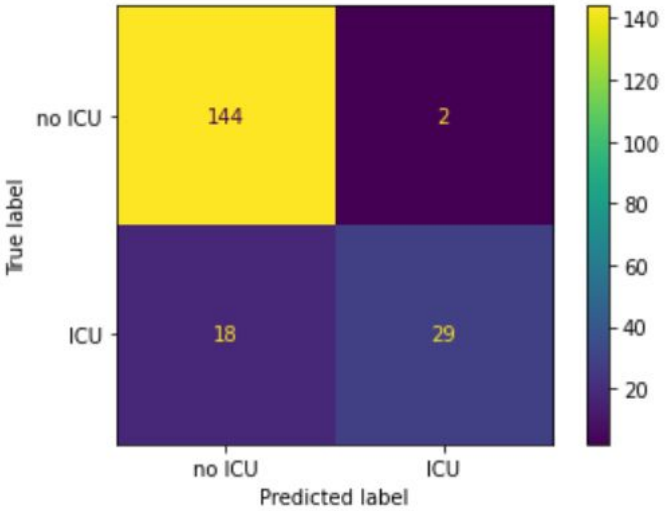

Figure 5. Error matrix using the random forest model on testing data.

Figure 5 shows the results from our error matrix on our testing data. The model correctly predicts 99% of the negative samples as negative. However, 38.3% of the patients that were actually admitted to the ICU was falsely predicted as not requiring ICU admission.

We looked at which features contribute most to our model's prediction to foster transparency in the model.

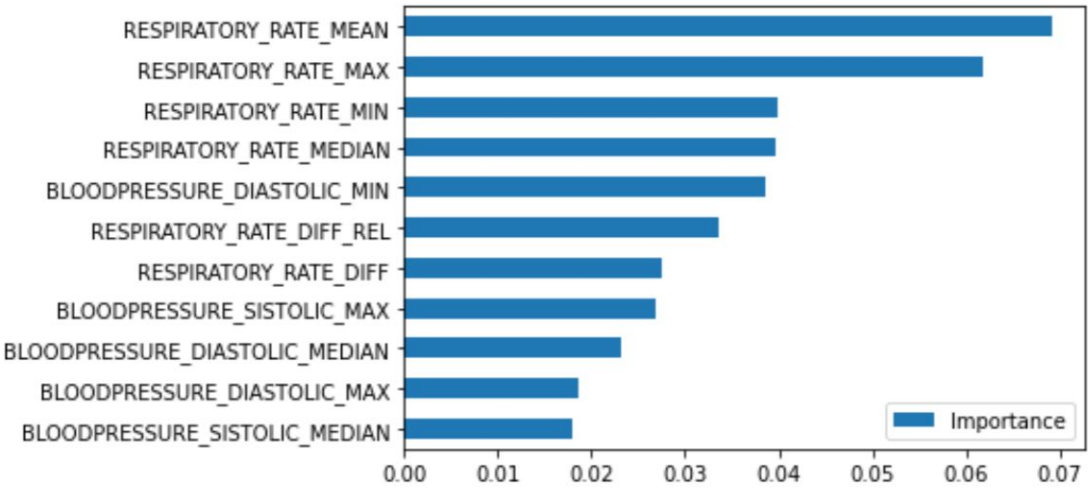

Figure 6. Feature importances for Random Forest Model

From Figure 6, we see that the respiratory rates as well as blood pressure measurements have the highest positive prediction coefficients. This means that higher respiratory rates and higher blood pressure measurements in a patient makes the model more likely to predict that the patient will be admitted to the ICU.

This is a very interesting finding as it aligns with previous clinical research stating that hypertension (high blood pressure) is associated with severity and mortality in Covid-19 patients. An increase in respiratory rate is also a likely sign of respiratory distress, which aligns with the well-known fact that covid-19 is a respiratory virus.

CAVEATS

Dealing with missing data

- There were 2 patients with missing vital sign values, hence a median value obtained from the training dataset was used. The missing data might have contributed to a different result had there been actual values in place of median values.

- *Low recall score*
  Our model has poor sensitivity. This is concerning because if our model was implemented in real life, resources for the ICU are underestimated as 2 out of 5 of patients requiring ICU admission are not predicted to be needing admission. This could be due to our unbalanced dataset. The model has more training data available for no admission prediction than training data for actual ICU admission.

Disease grouping 1-6 of patients not specified

- It is not known how the disease grouping may influence the model to predict admission to ICU. The model may give a different weightage to different but correlated disease groupings.

Window grouping of patients discrepancy

- It is unknown how long the patient had Covid-19 before being officially diagnosed. If the patient diagnosed with covid 19 had severe symptoms before undergoing covid-testing, it is highly likely the patient will be admitted to the ICU, thus this window is not normalized at baseline.

CONCLUSION AND FUTURE WORK

A poor recall score of 63% in our model shows the underestimation in the amount of admissions expected as only 3 out of 5 patients needing ICU facilities are predicted correctly. An acceptable model should be able to predict at least 9 out of 10 patients correctly as it is crucial that healthcare institutions and personnels are well prepared to save lives.

To build a more robust model, we should aim to improve the dataset we feed into our model. The current demographic information does not include geolocation of patients. Our data is merely from a few hospitals in Brazil and therefore is not generalizable on a global scale of places affected by covid-19. Future work should get more samples from diverse geologic and cultural locations.

REFERENCE

Dataset was obtained from https://www.kaggle.com/S%C3%ADrio-Libanes/covid19