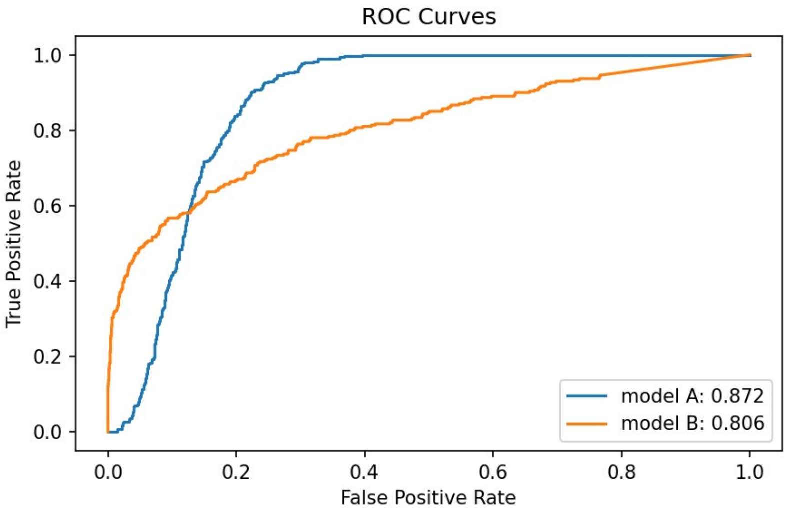
Phase 4: Model Evaluation

Welcome to Phase 4 of the capstone project. This section will be focused on the evaluation of the trained models. The questions will relate to the various challenges faced by the teams working on the two projects introduced in the first section.

You have made recommendations (based on your answers in the prior phase) to both of the research teams. They have taken your suggestions into account, and have since refined their results. Both teams have e-mailed you summaries of recent progress, which are shown below.

Project 1: CXR-based COVID-19 Detector

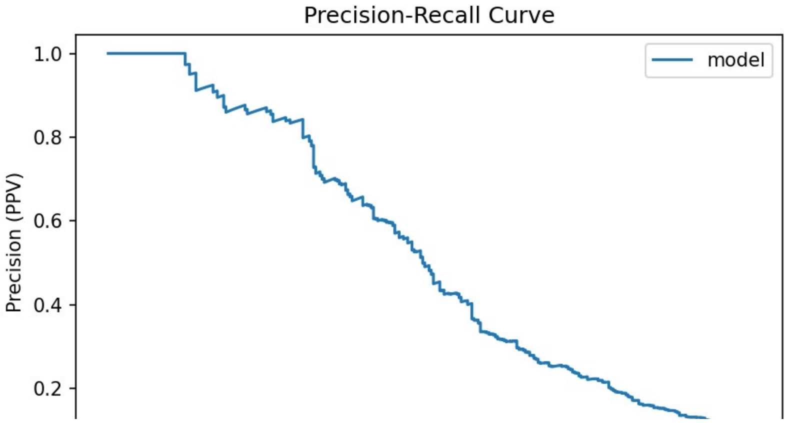
```
Hi,

Thank you for your excellent suggestions. We addressed the overfitting problem th
rough your excellent recommendations. We applied dropout with probability 0.5 and
added a small amount of random rotations to our data augmentation pipeline. We no
w have a couple of models, and we are interested in your opinion regarding the se
lection of the model with which we should proceed.

We trained models with early stopping twice- once using validation AUROC as the s
topping criterion and once using validation loss as the stopping criterion, and w
e plotted the ROC curve of both models.
```



```
The first model, model A, is the model chosen because it had the highest validati
on AUROC. The second model, model B, is the model chosen because it had the lowes
t validation loss. Let us know which model you think that we should use.

We understand that this is meant to be a COVID detector for the purposes of workl
ist prioritization. Could you help us pick a model given this use case? Let us kn
ow what you think, thank you.
```

Project 2: EHR-based Intubation Predictor

```
Hello!

We are happy to report that our new method of training on non-COVID samples has g
iven us strong results on the COVID dataset. Our results are very promising and w
e are excited to bring this project to its conclusion. In order to do that, we ne
ed to determine an operating point.

Given that this model will be used for automated triage, how should we reason abo
ut choosing a threshold? Below is the precision-recall curve for this model.
```



In [ ]: