Phase 2: Model Training, Part 1

Welcome to Phase 2 of the capstone project. This section will be the first of two parts that concerns the model training process of the model development cycle. You continue to play the role of a bioinformatics professor. The questions will relate to the various challenges faced by the teams working on the two projects introduced in the first section.

Your two research teams have begun working on the projects, and have some preliminary results. Both teams have e-mailed you summaries of progress thus far, which are shown below.

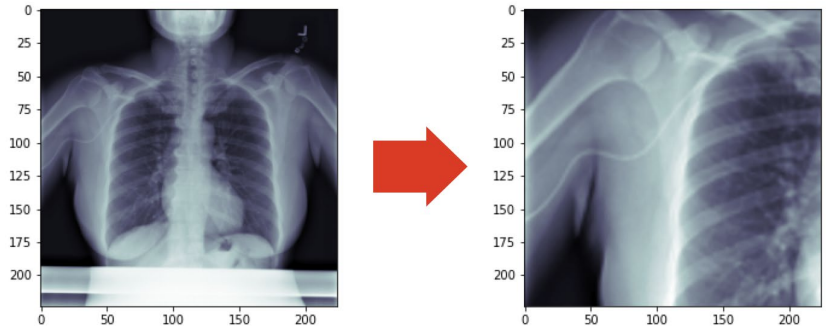
Project 1: CXR-based COVID-19 Detector

Hi,

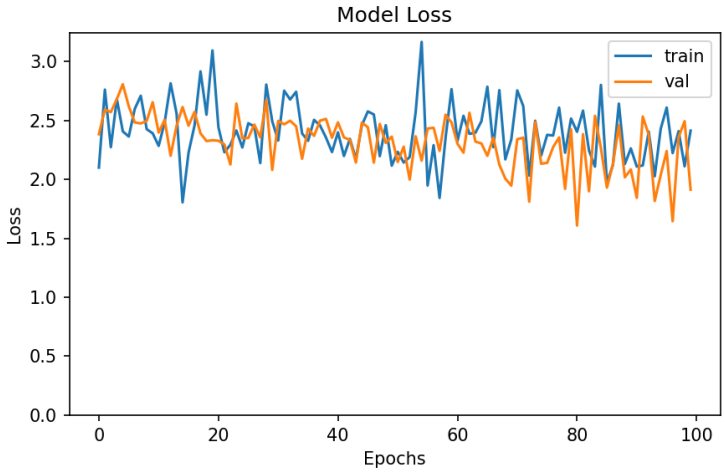
We are super excited to get this project kicked off! We have implemented the data pipeline and trained a few preliminary models, but there is still lots of room for improvement. Here is what we've done so far:

We split the data randomly into a training and test set. We are placing 90% of the data into the training set and 10% of the data into the test set. Additionally, the images were initially massive, on the order of 3000 by 3000 pixels. So, we re-sized the images to 224 by 224 pixels.

We are using the ResNet-50 CNN architecture. During training, we are applying data augmentation. Concretely, on a given image, with 50% probability, we are zooming in on a small, randomly selected region before feeding it to the model. Here is an example:



So far, we have seen the following training curves from our model. The loss for neither the training set nor the test set goes down very much.



As you can see, there is plenty of room for improvement. We'll keep working on it, but let us know if you have any suggestions. Thanks.

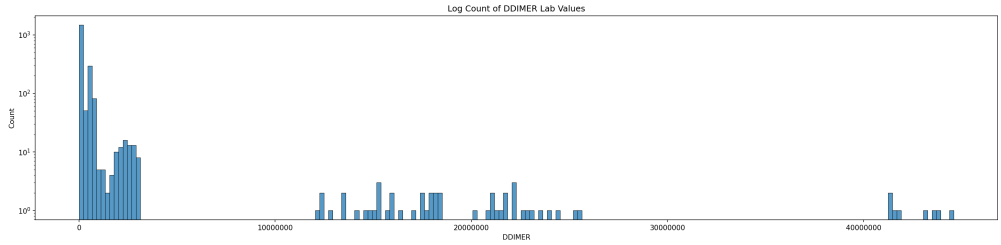
Project 2: EHR-based Intubation Predictor

Hello,

We are in the process of cleaning up the COVID EHR data, and expect to get a model training soon. We attempted to train a set of preliminary models, but ran into some data issues. We were wondering if you could take a look at some of the problems we've found in the data and let us know what you think.

First, we noticed that the EHR data is actually quite sparse relative to what we thought we have. We only have about 3,000 EHR records- not 30,000, as we originally thought. This leaves us with about 300 COVID-positive and 2,700 COVID-negative exams. We might not be able to train a model on this data alone.

We are noticing some very strange patterns in the data, particularly in the lab values. For example, see the following histogram of D-DIMER lab values found for each exam across the entire dataset. The x-axis is the D-DIMER lab values, and the y-axis is the number of exams with that count. We use a log-scale on the y-axis improve readability.



We saw this in several CSV columns, including Ferritin, and Procalci

tonin lab values. **We suspect that there is some underlying phenome na affecting all three lab values.**

Another issue we were running into were missing column values. We ca n't create a feature vector for Logistic Regression if we are missin g some values. How do you suggest we proceed regarding both the larg e outlier values and the missing values? Below is an example of the data once again, this time a sample of 30 exams (with the observed s ymptoms excluded). Note that NaN in the CSV means that the value is missing. Please take a look and let us know if you see something tha t we might have missed.

In [1]:
```
pd.read_csv('COVID_19_sample_data.csv')[
    ['pat_deid', 'intubation_date', 'IP_admission_date', 'IP_discharge_dat
e', 'clinic',
     'birth_date', 'death_date', 'gender', 'ethnicity', 'race_new', 'LYMAB
', 'CK', 'CR',
     'LDH', 'TNI', 'DDIMER', 'FERRITIN', 'PROCTL', 'PT', 'BUN', 'CRP',
     'SPO2', 'FIO2', 'NA']].iloc[5:35]
```

Out[1]:

	pat_deid	intubation_date	IP_admission_date	IP_discharge_date	clinic	birth_date	death_dat
5	8f9539f2-e6ad-4e00-ad45-4abc2bff2214	NaN	2020-03-04	2020-03-14	Clinic B	1965-11-11	Na
6	d5dd13c4-c31e-419c-8c02-47e4ca1ac5e2	NaN	2020-03-02	2020-03-20	Clinic B	2018-08-16	Na
7	91369e11-b944-4132-be0f-af46e880936b	NaN	2020-03-02	2020-03-21	Clinic C	1972-09-22	Na
8	c70992c9-ff13-467b-9032-1901506edeef	NaN	2020-02-29	2020-03-05	Clinic C	1959-06-17	2020-03-1
9	c70992c9-ff13-467b-9032-1901506edeef	2020-03-05	2020-03-05	2020-03-12	Clinic B	1959-06-17	2020-03-1
10	9ec7d743-96e7-47c8-b2ee-6336633beb39	NaN	2020-03-10	2020-03-23	Clinic B	1969-03-22	Na
11	9ec7d743-96e7-47c8-b2ee-6336633beb39	NaN	2020-03-23	2020-03-25	Clinic C	1969-03-22	Na
12	a527bcf0-3746-476c-	NaN	2020-03-03	2020-03-17	Clinic	1978-11-	Na

	90f2-dbab8868385e				C	27	
13	7f4ef129-1511-47a9-a9b7-8b0b2d02ad50	NaN	2020-03-12	2020-03-25	Clinic C	1952-05-06	Na
14	7078ae9a-4c79-4b30-b127-f76aabb6763e	NaN	2020-02-17	2020-03-07	Clinic B	1968-04-26	Na
15	a5c39700-6bf3-4984-af46-31344695e21b	NaN	2020-03-05	2020-03-13	Clinic A	1940-01-09	2020-03-1
16	a5c39700-6bf3-4984-af46-31344695e21b	2020-03-12	2020-03-12	2020-03-16	Clinic C	1940-01-09	2020-03-1
17	ddb2d5e2-643e-4374-ac19-f6ca3c0d16f5	NaN	2020-02-25	2020-03-09	Clinic C	1967-12-24	Na
18	21505aac-f219-43a8-ab3c-f57c6d8f1d1f	NaN	2020-03-08	2020-03-21	Clinic B	1940-05-03	Na
19	7992bf94-feee-4728-9187-2c911df2819b	NaN	2020-03-03	2020-03-17	Clinic C	2004-07-04	Na
20	d2f6d528-39db-4b7e-8389-abd27af9a710	NaN	2020-02-29	2020-03-12	Clinic B	1996-06-26	Na
21	fa0b58e6-6817-4d49-8211-1dd34abf0c15	NaN	2020-03-11	2020-03-28	Clinic C	2008-11-21	Na
22	b83237f3-9ff5-491e-aab4-d63ccff85f85	NaN	2020-03-13	2020-03-30	Clinic C	2012-11-17	Na
23	46988a9c-9c86-429a-bc4a-b3d14ff321b0	NaN	2020-03-11	2020-03-21	Clinic B	1957-03-13	Na
24	46988a9c-9c86-429a-bc4a-b3d14ff321b0	2020-03-20	2020-03-21	2020-03-24	Clinic B	1957-03-13	Na
	785b484d-7060-4d17-				Clinic	1942-08-	

25	bf18-ef8bbafc6f04	NaN	2020-02-28	2020-03-10	B	24	Na
26	edad31f3-5a08-4678-8d31-271a41a2aad5	NaN	2020-03-05	2020-03-13	Clinic C	1940-01-09	2020-03-1
27	edad31f3-5a08-4678-8d31-271a41a2aad5	2020-03-12	2020-03-12	2020-03-20	Clinic C	1940-01-09	2020-03-1
28	4607a669-4a97-4f0a-9661-856569905047	NaN	2020-03-09	2020-03-21	Clinic C	1993-11-26	Na
29	c1800ba1-7cba-45d7-bdc4-0e0b583932e4	NaN	2020-02-23	2020-03-08	Clinic A	2018-01-20	Na
30	d2718050-2e9c-4d5b-842e-52d910c1563f	NaN	2020-03-04	2020-03-17	Clinic C	1997-06-01	Na
31	d2718050-2e9c-4d5b-842e-52d910c1563f	NaN	2020-03-17	2020-03-22	Clinic A	1997-06-01	Na
32	818566cb-c89b-42d8-a6af-1a1ef13ed7cf	NaN	2020-03-08	2020-03-20	Clinic C	1984-10-11	Na
33	000e7adf-cbaa-4fad-ab2f-658c32f7d4d3	NaN	2020-03-12	2020-03-16	Clinic B	1959-01-03	2020-03-1
34	5a2f02ce-0286-45ae-b992-05331cb88379	NaN	2020-03-11	2020-03-29	Clinic C	1973-06-30	Na

In the following quiz, you will answer questions examining the issues of Team 1 and Team 2.

In [ ]: