COVID- 19

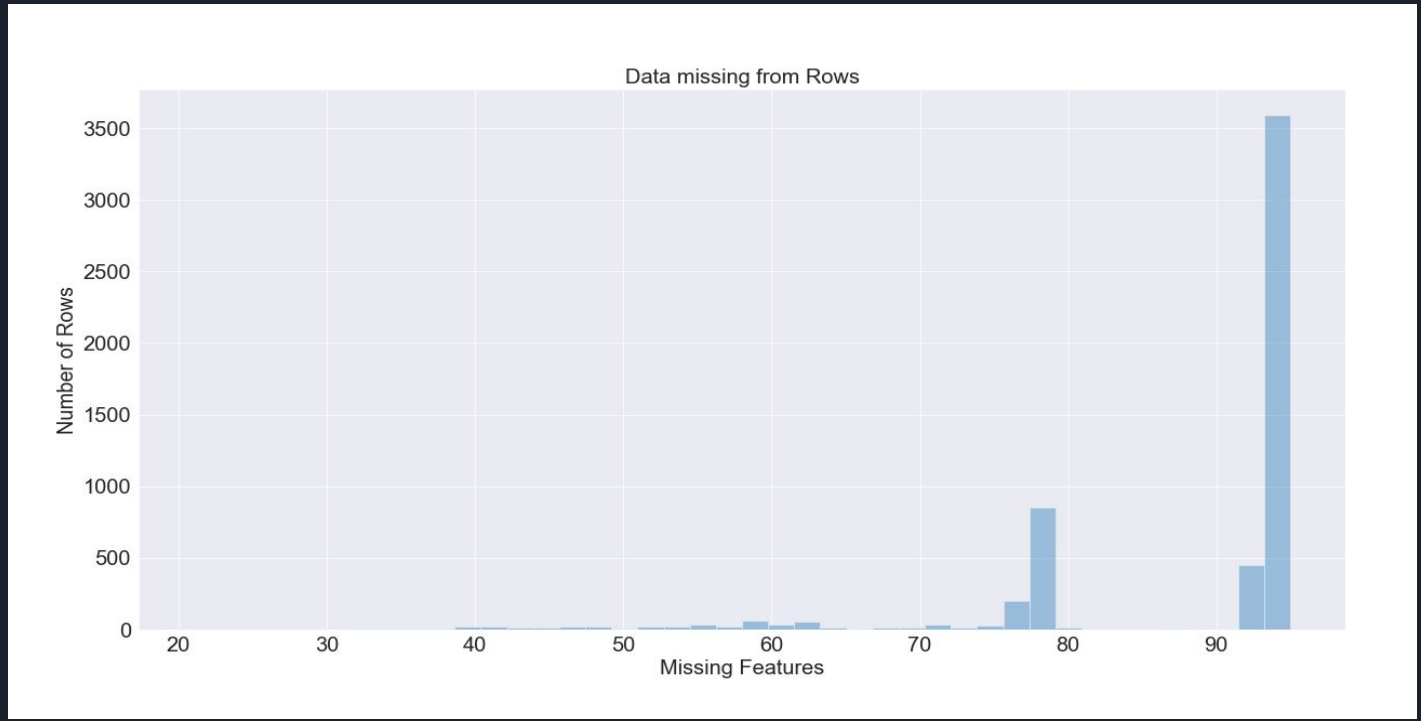
Predicting patient outcomes

The data and the objective.

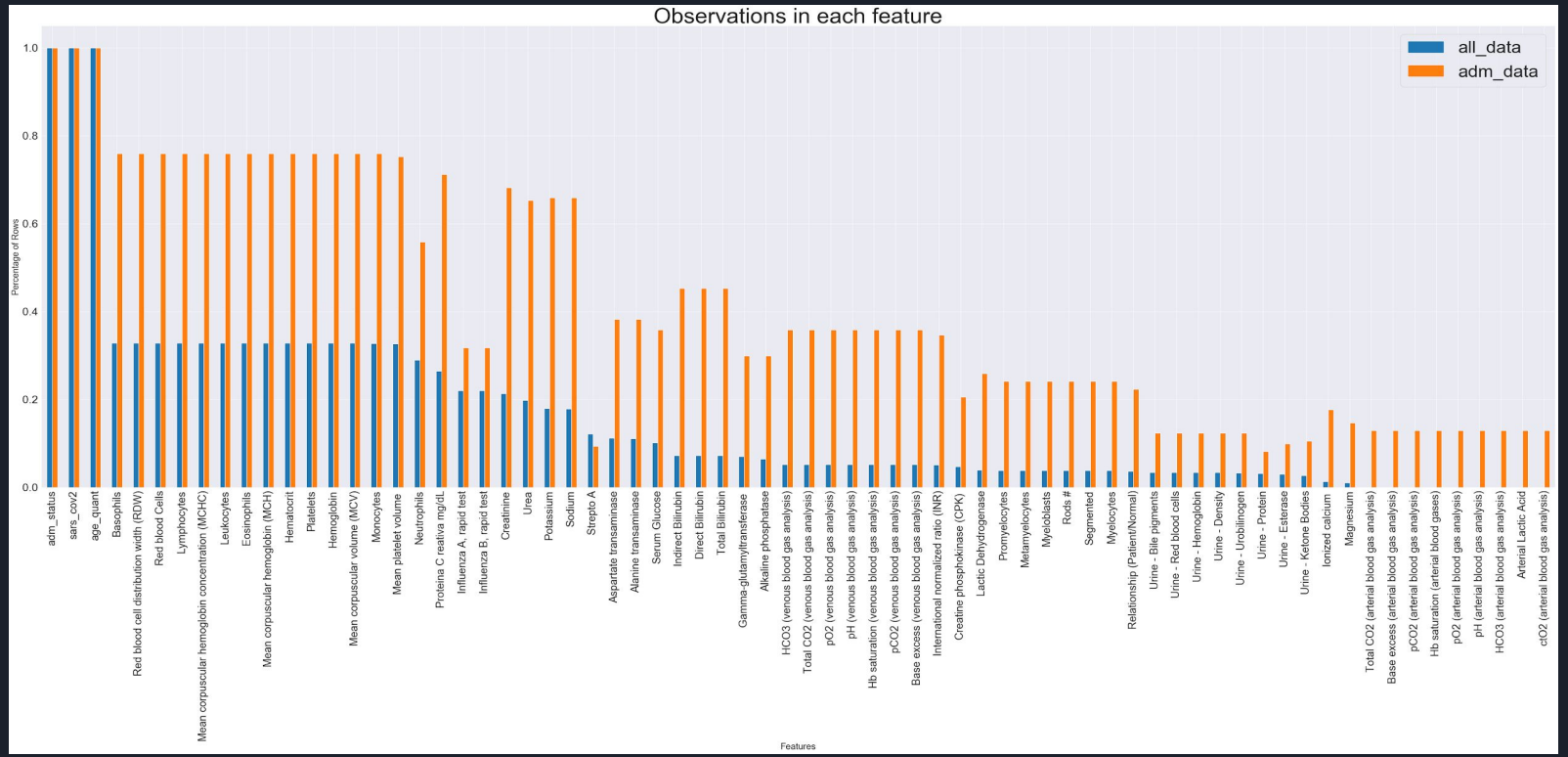
- Dataset contains anonymized data from patients seen at the Hospital Israelita Albert Einstein, at São Paulo, Brazil, and who had samples collected to perform the SARS-CoV-2 RT-PCR and additional laboratory tests during a visit to the hospital
- Based on the results of laboratory tests commonly collected among confirmed COVID-19 cases during a visit to the emergency room, would it be possible to predict which patients will need to be admitted to a general ward, semi-intensive unit or intensive care unit?
- Dataset was published on Kaggle and can be found at:

  https://www.kaggle.com/einsteindata4u/covid19

Dealing with missing values: Rows



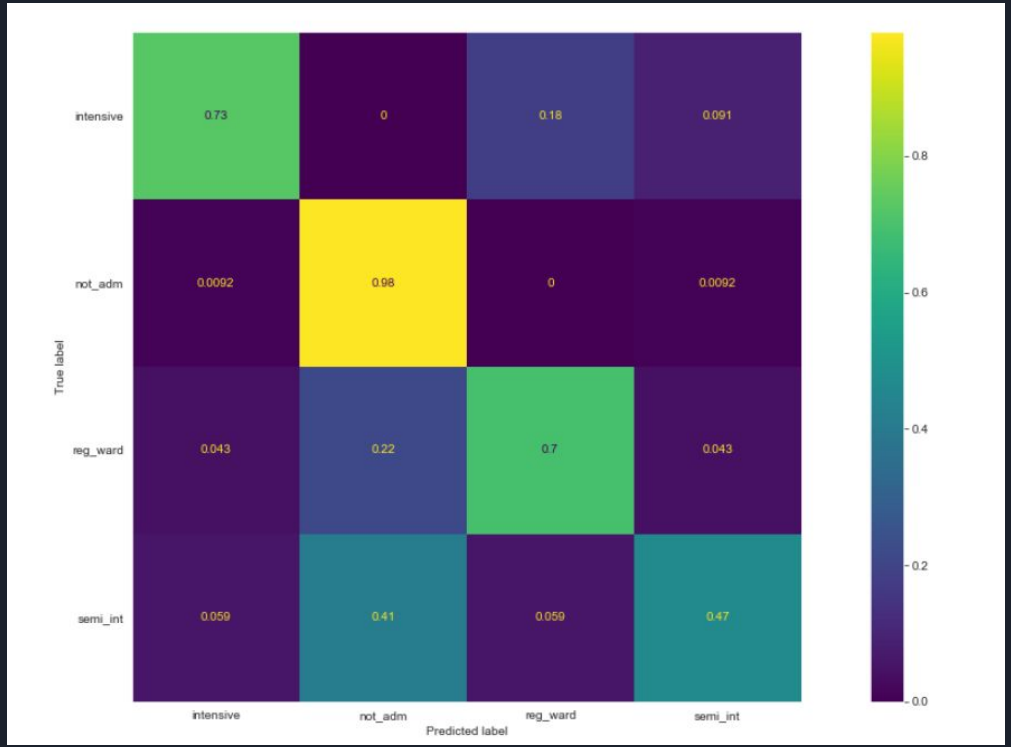
Dealing with missing values: Columns



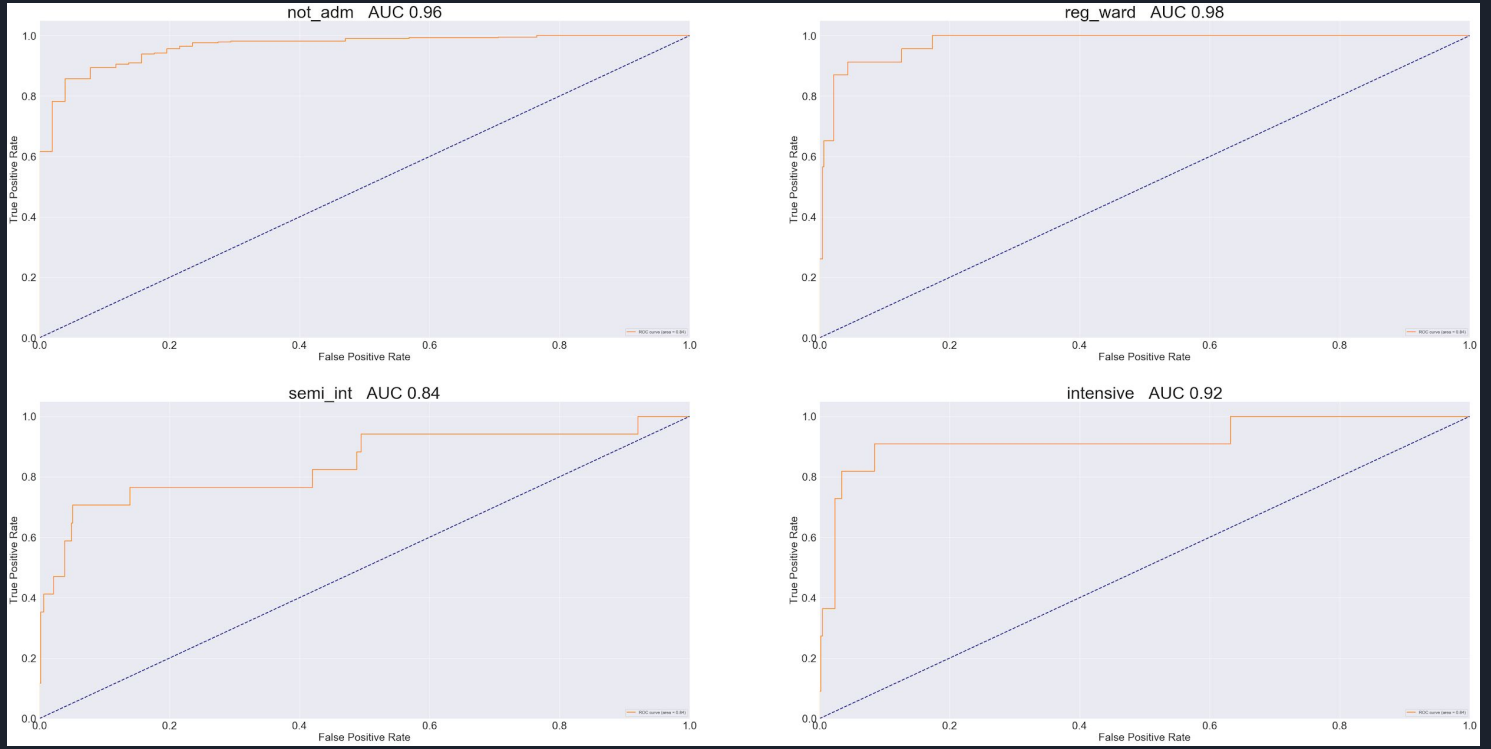
Modeling

- We ran multiple models and chose from those that performed the best.
- The tradeoffs were between precision vs recall.
  - How many people predicted as requiring intensive care actually needed intensive care?
    - **Vs**
  - How many people who actually needed intensive care were captured by the model?
- We prioritized recall, especially with respect to ‘intensive’
  - Its more important to identify people who need intensive care,
  - even if sometimes we accidently classify people who don't need intensive care as needing it.
- At the same we don't want precision to be too low.
  - We can make sure everyone who needs intensive care gets it by just sending everyone to intensive care
    - This is simply not possible and would be a huge burden on limited resources

Preliminary Model



Relabeling semi-intensive data



Model Performance.

	intensive_f1	intensive_recall	intensive_precision	reg_ward_f1	reg_ward_recall	reg_ward_precision
logis_f1_smote	0.857143	1.000000	0.750000	0.920000	0.958333	0.884615
logis_recall_dict_weight	0.814815	0.916667	0.733333	0.938776	0.958333	0.920000
logis_recall_smote	0.785714	0.916667	0.687500	0.920000	0.958333	0.884615
logis_f1_dict_weight	0.733333	0.916667	0.611111	0.916667	0.916667	0.916667
gs_svc_recall_smote	0.666667	0.833333	0.555556	0.875000	0.875000	0.875000
base_clf	0.640000	0.727273	0.571429	0.761905	0.695652	0.842105
base_semi_refilledf	0.620690	0.750000	0.529412	0.775510	0.791667	0.760000
gs_svc_f1_smote	0.583333	0.583333	0.583333	0.650000	0.541667	0.812500
xgboostt_f1_smote	0.411765	0.583333	0.318182	0.528302	0.583333	0.482759

Feature Importance