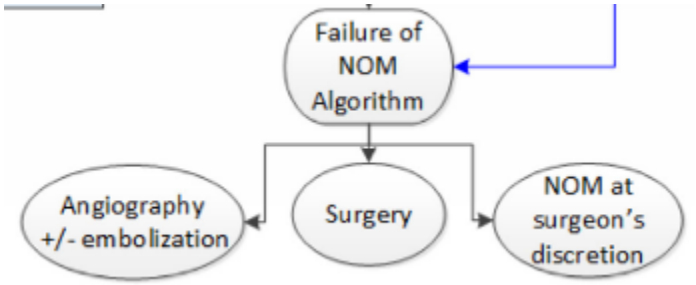
***Note**: All "**@@@_complete**" fields are largely made up of 2's, with some 1's and 0's. Not noted in the data dictionary.*
***Note**: Log skips over included features that didn't need notes.*
***Note**: Maybe if less than 5% of non-null data per column, don't use? (check all text-columns, "other" columns)*

CS180: Manual Feature Selection List

Removed: **id, redcap_data_access_group**

Include: *****time_to_angio***:** might be really significant but not noted in the data dictionary, indicates "failure of NOM Algo" aka an outcome to refer to, lots of NULL - *must ask if NULL means no intervention or just no angio.* **Note:** there is an "**angiography**" and "**embolization**" feature



Removed: **mo_injury, yr_injury, ins**

Unsure of significance: **demographics_complete**

Include: **hx_trauma and hxtrauma_days:** can be condensed into 1 column since hxtrauma_days has a range of 1-364. can just use 0 for the other values, hxtrauma_days only has about 4 non-null numbers out of the 1000 patients. so we should just use hx_trauma as a true/false feature between known and unknown trauma within the last year OR between definitively no trauma and no known trauma.

Include: **osh_used:** only a couple unknowns (888), can exclude those patients?, then only use 0-No 1-Yes.

Unsure if we should use: **time_ed, time_ptc** (timely response rate might matter)

Include: **moi:** method of injury, one-hot encoding needed for 11 categories +1 unknown (888), but decent amount of "others" noted in text in **moi_other**

Include: **bleed:** leave as is, only has 0/1 values for no/yes.

Maybe remove: **admit** (lots of non-null data, but not sure if necessary)

Unsure of significance: **injury_characteristics_and_admission_complete**

Include: **unstable_elem_#:** #1-#9 and 0 for None, lots of data noted in text in **other_elem** like moi

Unsure if we should use: **ems_used** (lots of non-null data, but not sure if necessary)

Include: **sbp_field, dbp_field, sbp_lowed, dbp_lowed, sbp_lowptc, dbp_lowptc, hr_lowfield, hr_highfield, hr_lowed, hr_highed, hr_lowptc, hr_highptc, hypo_ed**, hypo_aftered**, vaso_first48, vasos_# (1-7 w/ 888), vaso_other*****

- Deal with many NULL values (can try doing average of columns?)
- ed = emergency dept., ptc = pediatric trauma center
- ** maybe need to combine/manipulate these two, mostly 0's but there is a decent amount of data
- *** can likely get rid of vaso_other because there's only ~10 non-blank

Unsure if we should use: **vaso_hrs** (little data, might not be enough to use)

Include: **intubated_field, intubated_ed, which_ed__# (1,2,888)**

Unsure of significance: **vitals_complete**

Include: **add_injuries__# (1-10), gcs, iss, ais_head, ais_chest, ais_face, ais_abdomen, ais_extremity, ais_external**

- Deal with many NULL values (unsure what method to use)

Unsure of significance: **scores_complete**

Include: **hb# and time#:** 1-37, will standardize into 1 column for each interval of hours

Unsure of significance: **labs_complete**

Include: **transfusion, mtp, transfusion_hrs**, transfusion_local*****

- **only used if transfusion == 1 (unsure of how to handle remaining nulls)
- ***only used if transfusion == 1 and is categorical, may need to separate

Unsure of significance: **prbc_24osh, prbc_osh, prbc_24ptc, prbc_ptc, ffp_24osh, ffp_osh, ffp_24ptc, ffp_ptc, platelets_24osh, platelets_osh, platelets_24ptc, platelets_ptc, blood_products_complete**

- Because of LOTS of NULL data

Include: **fast and fast_results:** results (1-3 w/ 888) rely on fast (0/1 values w/ 888). Maybe combine into 1 column

Include: **spleen_grade, spleen_fluid, spleen_ctblush, liver_grade, liver_fluid, liver_ctblush, kidney_grade*, pancreas_grade***

- Data is NULL if injury is only in spleen/only in liver, not NULL for both if both injured
- *kidney and pancreas grade largely 0's but there is some data

Remove: **embo_reason** (largely NULL, only has values if there was **embolization**, but even then there's so little data)

Remove: **Kelly's thought: the rest of embo: **embo_type, embo_other, embo_tech, embo_techother,** is not currently important enough to include. If we already know they did a non-NOM on the patient, it's all we really need to know for now. If we have time to include/manipulate, maybe include.

- But they separated these features into multiple columns for us already?
- Largely 0's anyway

Unsure of significance: **radiology_complete**

Maybe include (likely not): **dpl, exlap, exlap_hrs, exlap_damage, lap, lap_hrs, lap_open, exlap_lap_purpose**

- Largely 0, expands into **exlap_hrs, exlap_damage** with is largely NULL

Include: **procedure_# (0-12), procedure_other***

- already separated into columns for each category for us
- *other category is in text, (unsure of how to handle, maybe remove)

Maybe remove: **percu_drain, percu_hrs, percu_fluid__# (0-7, 888), percu_other, percu_performed, percu_performed_other, percu_us**

- Largely 0 or NULL. ~11 data points, maybe not enough data to be significant, even though the feature description seems useful

Include: **other_procedure**
Remove: **other_procedure_describe**

- Might be too hard to use the "describe" text. But there is decent amount of **other_procedure** data

Unsure of significance: **procedures_complete**

Include: **missed_iai** (seems like a significant feature through description, even if little data)
Maybe include: **describe_iai** (could be hard to use, and there's little data)

Remove: **missed_ii, describe_ii** (basically no data, only like 3 datapoints)

Include: **missed_any** (seems like a significant feature through description, even if little data)
Remove: **describe_any** (could be hard to use, and there's little data)

Remove: **missed_treatment** (not very relevant/not much data)
Include: **describe_treatment** (very little data, but keywords of "surgery")

Include: **hosp_comps and comps__#:** comps (1-14) is separated into columns already, hosp_comps can be converted to a "none" column for comps (invert 0's and 1's)

Maybe remove: **ards_vent** (not a lot of data)**, reor_operation** (not a lot of data), **trans_reaction** (all NULL)

Include: **comps_other** (lots of data, but how to handle all text?)

Include: **infection__# (0-6, 888), infection_other** (lots of data, but how to handle all text?)

Include: **disposition** (3 categories, might separate into 3 columns, unless not a lot of "3's")

Include: **cause_death** (mainly 3's and 6's, maybe can make use of)**, cause_other** (mostly bleeding or brain injuries, can manually manipulate)

Unsure of significance: **outcomes_complete**

Remove: **pt_status** (redundant)
Include: ALL **stable_** and **unstable_**, if NULL, then change to 0 because it's mutually exclusive
Remove: **stable_deviation_decscribe** (could be unnecessary and very very little data)

Include: **unstable_notprotocol_other** (a lot of "death" text, may need to manually comb through for consistency)

Unsure of significance: **algorithm_adherence_complete**

~~*Remove:* **fu1_intervention_procedure, fu1_radiology, fu1_radabd, fu1_abdct, fu1_abdct_finding__# (1-4), fu1_abdct_other** (largely NULL or 0's)~~

Decision: disregard all follow-ups (**fu1_ and fu2_**). Doesn't really solve the problem at hand, whether the patient is actively in shock and needs surgery. fu#_intervention and fu#_intervention_procedure not actually necessary.

Include: **hgt, bmi, firsttemp_ptc, lowtemp_ptc,** (~50% NULLs and 888s, but must deal with)

Remove: **firsttemp_ptc_method, lowtemp_ptc_method, lasttemp_ptc, lasttemp_ptc_method, firsttemp_icu, firsttemp_icu_method, time_normtemp**

- Just so much missing data, mostly normal temperatures/similar temperatures to firsttemp_ptc and lowtemp_ptc

Unsure of significance: **additions_complete**

Note: *Rest of columns in the dataset seems to be used for their own KNN prediction. May not be relevant to ours, or not enough information to be using those columns.*