COMP40350: HW1: Data Quality Report

Adam Ryan - 14395076

March 12, 2021

Contents

1 Background

COVID-19 is an infectious disease caused by SARS-CoV-2, a coronavirus strain discovered in December 2019 first identified following an outbreak in the Chinese city Wuhan, with the WHO declaring the outbreak a global pandemic in March 2020.

Since its discovery, health organisations have been actively gathering data to assess aspects of the disease including infectivity, symptoms, and mortality rate. Active interest has been paid to factors which may increase a patient's risk of serious symptons or death.

In this analysis, we focus on using the data collected by CDC to build an analytics solution for predicting a patients' death risk prediction. CDC collects demographic characteristics, exposure history, disease severity indicators and outcomes, clinical data, laboratory diagnostic test results, and comorbidities. It also includes information on whether the individual survived or not.

2 Overview

This report will outline the initial findings based on the provided sample of the CDC dataset. It will summarise the data, describe the various data quality issues observed and how they will be addressed.

Appendix includes includes feature summaries and boxplots used to visualise the data (featured in notebook and containing folder)

3 Summary

The following are the key points in relation to the data set and approach:

- The dataset lacks a primary key.
- The dataset lacks a patient identifier so we cannot look for readmitted patients.
- The dataset consists of 10,000 rows and 12 (non-repeated) columns.
  - 4 features are datetime features formatted as YYYY/MM/DD.
  - 3 of these are demographic
  - 4 are booleans with additional values
  - 1 is a diagnosis classifier
  - All features are categorical in nature, however there is scope for the addition of featuers.
- While null values are largely absent from the dataset, there are high proportions of values flagged as 'missing' and 'unknown', with some features containing both missing and unknown values. The distinction between missing and unknown should be confirmed with a source knowledgeable on the data set, however the author's initial recommendation is that these features would be likely targets for imputation mapping both features to a single 'unknown' value.
- The datetime columns are most heavily affected by null or missing values. The author notes that the cdc data dictionary highlights the depreciation of the cdc_report_dt column and points to the usage of cdc_case_earliest_dt in its place. Following this, the author recommends removing the now depreciated cdc_report_dt column.
- 515 rows feature a cdc_case_earliest_dt date which does not match the minimum of the other datetime columns as implied in the cdc data dictionary. These instances should be checked with a domain expert to understand the source of these dates and if they are valid. The author suggests including these until later revision.
- The volume of duplicate rows is low at 431 rows (4.3%). Investigation into the cause of duplicates highlights sparse population of data, or common data population, is the primary cause of duplicates (e.g. racial info is missing in 90% of duplicate instances, with icu and medical condition info missing in over 95% of duplicate instances. Although these instances are likely 'valid', the recommendation is to drop these duplicate instances as the high prevalence of missing information is unlikely to provide useful information into our model.
- The features prevalent are good targets for conversion to a 'category' datatype with limited valid values prevelant across all category features.
- There is one record where there is an icu admission flagged but not a hospital admission. This record should be removed due to inconsistency in the data and low impact on overall set.
- The current_status column contains 93% Laboratory confirmed cases. It should be identified with a domain expert as to whether the probable cases must be considered. If the probable cases can be dropped, the recommendation is to remove the probable cases and remove this feature however this will be included in further components of the analysis.

4 Logical Integrity

As the dataset has a heavy focus on categorical data, the following tests were carried out to asses the integrity of the dataset

- T1: Check if there are cdc_case_earliest_dt's which are not the earliest of the other dates.
  - Result: 2857 (29%) Records which are not the earliest
  - Result: 515 (5%) Records which are not the earliest where not all of the other dates are populated
  - Query: Where does this data come from?
- T2: Check if there are ICU admissions without hospital admissions.
  - Result: 1 Record which should be updated
- T3: Check if there are probable cases with a confirmed positive specimen
  - Result: 227 Record which should be updated to laboratory confirmed
  - Result: 248 Records when hospital admission is also true.

5 NonDatetime Categorical Features

There are 8 noncategorical features in the dataset:

- F1: current_status - A feature to flag if the case is confirmed via lab or suspected.
  - Null: Not applicable.
  - Top Value: Laboratory Confirmed Case - 93% of rows.
  - Unique Values: 2
  - Overall data is reasonable. Actionable item to update probable cases where there is a positive lab specimen.
- F2: sex - A feature to flag the patients' sex.
  - Null: Not applicable.
  - Top Value: Female 53%
  - Unique Values: 4
  - Should be updated to combine unknown values
- F3: age_group - A feature to flag the patients' age group.
  - Null: Not applicable.
  - Top Value: 20-29 Years 18%
  - Unique Values: 10
  - 14 records have an unknown age grouping.
- F4: race_ethnicity_combined - A feature to flag the patients' ethnicity.
  - Null: Not applicable.
  - Top Value: Unknown 41%
  - Unique Values: 10
  - 41% unknown values.
  - Contatenated field with comma separated values. Separation denotes Hispanic or not. This info is already captured via the racial component.
- F5: hosp_yn - A feature to flag if the patient was hospitalised.
  - Null: Not applicable.
  - Top Value: No 52
  - Unique Values: 5
  - Missing and unknown two separate values. OTH present in one record.
- F6: icu_yn - A feature to flag if the patient was admitted to ICU
  - Null: Not applicable.
  - Top Value: Missing 77%

  - Unique Values: 4
  - Check with domain expecrt on missing % reason. Are missing values indicative that the patient never ended up in the ICU and hence it was not flagged? Initial investigation suggsts that Missing Corresponds with No. Note in particular that for younger patients are more heavily represented as a percentage of their age group within the missing category, and similarly older patients are more likely to be represented in the no category than younger patients as a proportion of their age group (something that would appear contradictory). My initial recommendation would be to populate this with 'no' where it is missing, however I would leave any population of the value as the final actionable step so that the ML model can be easily tested with and without this value to decide on a sensible approach. I suspect that older patients are flagged explicitly as being non-ICU patients as there might be more concern over it being needed resulting in an almost skewing of the value.

- F7: medcond_yn - A feature to flag if the patient had comorbidities.
  - Null: Not applicable.
  - Top Value: Missing 75%
  - Unique Values: 4
  - 82% unknown and missing values.

- **Target Feature**: death_yn - A feature to flag if the patient died.
  - Populated. 3% are yes.

6 Datetime Categorical Features

There are 4 categorical datetime features in the dataset:

- D1: cdc_case_earliest_dt - A feature to flag if the case is confirmed via lab or suspected.
  - Null: Not applicable.
  - Top Value: Laboratory Confirmed Case - 93% of rows.
  - Unique Values: 2
  - Overall data is reasonable. Actionable item to update probable cases where there is a positive lab specimen.
  - Covers 325 2nd January 2020 to 16th January 2021 (missing days present)
- D2: cdc_report_dt - A depreciated column. CDC recommendation is to drop for D1.
  - Should be dropped due to depreciation.
- D3: pos_spec_dt - First positive specimen collected
  - Null: Yes 72% missing
  - Rec: Use to update Status and drop as missing percentage too high.
- D4: onset_dt - Date of symptom onset
  - Null: Yes 49% missing.
  - Top Value: Unknown 41%
  - Unique Values: 326
  - 41% unknown values.
  - Keep for determining time between reporting and symptom onset.
  - Covers 2nd January 2020 to 28th January 2021.

7 BoxPlots

BoxPlots were produced for all categorical data. These are present in the appendix due to the size of the file. All pairs of data and single value info was calculated as an initial exploration. Initial exploration consists of files beginning with prefix 01 and 02 respectively.

8 Appendix

Please see containing folder for files generated visualising feature relation.

9 Note

The steps provided in the assignment outline more of a linearisation in the process, however upon reviewing the data I did not believe the outlined processed was particularly suitable for this dataset.

In particular, the processing steps outlined suggest the removal of duplicate values prior to data exploration. As I did not beleive the records were, in fact, duplicates but instead were driven by other elements, it was more reasonable to explore the relationships between various factors before taking any steps to drop rows with overlap, in order to better understand why.

Similarly, the steps provided suggest not adding columns until the final section. Due to the nature of the data and the variety of missing values within some of the indicator and date columns, it seemed to me that valuable information could be obtained based on my initial exploration before any final removal occurs. In particular, the onset datetime column looks to have key value in relation to the asymptomatic prevalence of COVID and the time between initial presentation and symptom onset date. Therefore, adjusting the nature of this column and adding on attributes which reflected the data that was in the original column while preserving and enhancing the data set was logical as an approach before simply dropping this feature for missing prevalency. Similarly, the race column contains race and ethnicity combined however this can be replaced with the racial info as that alone is sufficient to capture the concatenated nature of this. While there may be a need from a reporting purpose in the CDC to compare Hispanic vs Non-Hispanics demographics, reducing the memory usage of the field by stripping the redundant info still allows recovery if this would be insightful.

Due to all of the above, the data quality plan and data quality actioning were, in a sense, completed as a joint process as proper cleansing of the set did not allow for a full linearisation of this process. This steps is detailed below.