Machine Learning Project using EHR Data

Capstone Project
Data Science Batch of 8-Feburary 2020

Divyansh Chahar

https://www.linkedin.com/in/divyanshchahar/

https://github.com/divyanshchahar

Thursday 4th March, 2021

1 Introduction

Introduction goes here.

2 EDA approach

EDA for this model was divided into 2 main parts i.e. understanding the entries in a particular file and understanding the insights.

2.1 Understanding file entries

2.1.1 allergies.csv

File	Operation
allergies_uvc1.csv	Unique Values Count on "DESCRIPTION"
allergies_uvc2.csv	Unique Values Count on "PATIENT"
allergies_nuvc1.csv	Unique Values Count on "PATIENT" on dataframe filtered by null values in "STOP"
allergies_unwanted.csv	Count of unwanted data in the dataframe

Table 1: Files with results of EDA on allergies.csv

- **Allergy to mould** is the most frequent entry and **Allergy to soya** is the least common entry (see *allergies_uvc1.csv*)
- It is possible for a single patient id to be listed against multiple enteries in *allergies.csv*, (*see allergies_uvc2.csv*) however is it possible for a patient to have multiple allergies remain to be seen.
- It is possible for a patient to have multiple ongoing allergies (see *allergies_nuvc1*)
- "STOP" column in *allergies.csv* have some null values, but this cannot be attributed to unwanted or missing data, it represents entries with ongoing allergies(see *allergies_unwanted.csv*).

2.1.2 careplans.csv

File	Operation
careplans_uvc1.csv	Unique Values Count on "PATIENT"
careplans_uvc2.csv	Unique Values Count on "DESCRIPTION"
careplans_uvc3.csv	Unique Values Count on "REASONDESCRIPTION"
careplans_nuvc1.csv	Unique Values Count on "PATIENT" on dataframe filtered by null values in "STOP"
careplans_nuvc2.csv	Unique Values Count on "DESCRIPTION" on dataframe filtered by null values in "REASONDESCRIPTION"
careplans_unwanted.csv	Count of unwanted data in the dataframe

Table 2: Files with results of EDA on careplans.csv

- It is possible for a patient to have multiple careplans (see *careplans_uvc1.csv*), however further investigation is required to check if it is possible for a patient to have multiple ongoing careplans.

- It is possible for a single patient to have multiple ongoing careplans (see *careplans_nuvc1.csv*)
- To study the most common reasons careplans are used for, it was deterimined to perform a unique Values count on the columns named "DESCRIPTION" and "REASONDESCRIPTION" of the *careplans.csv*. Perfomring a unique values count on either one of the column will not be sufficient as single value in the "DESCRIPTION" columnn could be listed against multiple entries in the "REASONDESCRIPTION" column.
- Careplans are most commonly used for **respiratory therapy** (see *careplans_uvc2.csv*)
- Careplans are most commonly used for treatment of **Acute Bronchitis(Disorder)** (see *careplans_uvc3.csv*) which is a respiratory disorder thus justifying the fact that Respiratory Therepy is the most commonly occuring value in the "DESCRIPTION" column
- The "STOP" column contains NaN values but it could be attributed to ongoing careplans and cannot be considered as unwanted or bad data. The "REASONDECSRIPTION" column also contains NaN values it represents the use of careplans for non-medical reasons. Thus we can say that *careplans.csv* does not contains any unwanted data (see *careplans_unwanted.csv*).
- For non-medical reason (i.e. when "REASONDESCRIPTION" has a NaN value) the careplans are most commonly used for **Self-care interventions (procedure)** (see *careplans_nuvc2.csv*).

2.1.3 conditions.csv

File	Operation
conditions_uvc1.csv	Unique Values Count on "DESCRIPTION"
conditions_nuvc1.csv	Unique Values Count on "DESCRIPTION" on dataframe filtered by null values in "STOP"
conditions_psudo_chronic.csv	List of psudo chronic conditions i.e conditions which have no "STOP" values in some cases but do have a "STOP" value somewhere else.

Table 3: Files with results of EDA on conditions.csv

- Viral Sinusitis (disorder) was the most common condition (see *conditions_uvc1.csv*)
- It was observed that the "STOP" column in the *careplans.csv* had some null values, thus giving the impression that certain conditions could be chronic in nature thus the dataset was filtered by null values in the "STOP" column and unique values count was performed on the "DESCRIPTION" column, the results of this operation are stored in *condition_nuvc1.csv*. However this approach has a major drawback, some conditions which do not have a stop date in certain cases do have stop dates in some other cases.
- In order to check weather a condition is trully chronic in nature or not, further investigation was required. The dataset was checked to see if any of the conditions in the *conditions_nuvc1.csv* has stop dates anywhere in the dataset. The enteries pertaining to this investigation are recoded in *conditions_psudo_chronic.csv*.
- To get the list of final chronic conditions, a difference operation was performed between the list of conditions in *conditions_STOP_null.csv* and *conditions_nuvc1.csv*. The resultant conditions were recorded in *conditions_chronic.csv* stored in the *preprocessed* folder

2.1.4 encounters

File	Operation
encounters_uvc1.csv	Unique Values Count on "REASONDESCRIPTION"
encounters_uvc2.csv	Unique Values Count on "DESCRIPTION"
encounters_uvc3.csv	Unique Values Count on "ENCOUNTERCLASS"
encounters_nuvc1.csv	Unique Values Count on "DESCRIPTION" filtered by NaN values "REASONDESCRIPTION"
encounters_nuvc2.csv	Unique Values Count on "ENCOUNTERCLASS" filtered by NaN values "REASONDESCRIPTION"
encounters_ct1.csv	Cross Tabulation operation between "REASONDESCRIPTION" and "ENCOUNTERCLASS"
encounters_ct2.csv	Cross Tabulation operation between "DESCRIPTION" and "ENCOUNTERCLASS"
encounters_ct3.csv	Cross Tabulation between "DESCRIPTION" and "REASONDESCRIPTION" column
encounters_unwanted. csv	Count of Unwanted data on the dataframe

Table 4: Files with results of EDA on encounters.csv

- Most encounters happen because of non-medical reasons i.e. **NaN values** and the second most common cause of encounters is **Normal Pregenancy** (see *encounters_uvc1.csv*).
- As discussd above the most encounters are due to non-medical reasons, a further analysis reveals that most encounters can be classified as **Well child visit (procedure)** (see *encounters_uvc2.csv*) .
- For most of the encounters the patient was in an ambulatory state i.e. the patient is not bed ridden and can walk (see *encounters_uvc3.csv*).
- *encounters.csv* was also analyzed for unwanted data and the output was stored in *encounters_unwanted.csv*. It was observed that only "REASONDESCRIPTION" had **NaN values** and "PAYER_COVERAGE" had some zeros. It is possible that a patient had no insurance coverage at all hence the zeros in the "PAYER_COVERAGE" column cannot be treated as unwanted values values however the null values in "REASONDESCRIPTION" column need to be investigated further. Thus at this point it can't be stated with absolute certainty that *encounters.csv* does not have any unwanted data (see *encounters_unwanted.csv*).
- For encounters which are not because any major medical purpouse the most common enounters are for **Well child visit (procedure)** (see *encounters_nuvc2.csv*) . After comparing the data in *encounters_uvc2.csv* and *encounters_nuvc1.csv* we can say that all the encounters of the type **Well child visit (procedure)** are for non-medical reasons.
- It was also observed that most of the encounters for non-medical reason are most commonly of the type **wellness** (see *encounters_nuvc2.csv*)
- On a closer observation it was noticed that a single entry in "DESCRIPTION" column could be listed against multiple entries in "REASONDESCRIPTION" and "ENCOUNTERCLASS". Thus several cross tabulation operations were performed to understand the distribution of the data across multiple categories.

- Most of the entries in "REASONDESCRIPTION" and "DESCRIPTION" could be listed across multiple entries of "ENCOULNTERCLASS" (see *encounters_ct1.csv* and *encounters_ct2.csv*).
- It must also be noted that several entries in the "DESCRIPTION" column could be listed against more than one entry in the "REASONDESCRIPTION" column (see *encounters_ct3.csv*) .
- The "ENCOUTERCLASS" column has several unique values, we are particularly interested in the encounters where "ENCOUTERCLASS" is **emergency** and **urgentcare**.

2.1.5 imaging_studies.csv

File	Operation
imaging_studies_uvc1.csv	Unique Values Count on "BODYSTIE_DECSRIPTION"
imaging_studies_uvc2.csv	Unique Values Count on "MODALITY_DESCRIPTION"
imaging_studies_ct1.csv	Cross Tabulation operation on "BODYSITE_DESCRIPTION" and "SOP_DESCRIPTION"
imaging_studies_ct2.csv	Cross Tabulation operation on "MODALITY_DESCRIPTION" and "SOP_DESCRIPTION"
imaging_studies_unwanted.csv	Count of Unwanted Data in Dataframe

Table 5: Files with results of EDA on imaging_studies.csv

- Thoracic Structutre (Body Study) is the most examined body structure (see *imaging_studies_uvc1.csv*) .
- **Digital X-ray** is the most commonly performed procedure (see *imaging_studies_uvc2.csv*).
- It can be noted that multiple tests could be performed on a single body part.
- *Thoracic structure (body structure)* is the only value in "BODYSITE_DECSRIPTION" that is listed against multiple values in the "SOP_DESCRIPTION" column (see *imaging_studies_ct1.csv*) . It must also be noted that "BODYSITE_DESCRIPTION" also have values **Thoracic structure** and **thoracic** which are listed against a single value in "SOP_DESCRIPTION" column. However if these represent different body site or are just a different values for the the same body site is a topic of further analysis. All ohter values in "BODYSITE_DESCRIPTION" column are listed against a single value in "SOP_DESCRIPTION" column.
- *imaging_studies.csv* does not contains any unwanted data (see *imaging_studies_unwanted.csv*)

2.1.6 immunizations.csv

File	Operation
immunization_uvc1.csv	Unique Values Count on "DESCRIPTION" column
immunization_unwanted.csv	Count of Unwanted Data on Dataframe

Table 6: Files with results of EDA on immunizations.csv

- **Influenza seasonal injectable preservative free** was the most frequently occurring entry (see *immunization_uvc1.csv*).

- *immunization.csv* contains no unwanted data (see *immunization_unwanted.csv*)

2.1.7 medications.csv

File	Operation
medications_uvc1.csv	Unique Values Count on "DESCRIPTION" column
medications_uvc2.csv	Unique Values Count on "REASONDESCRIPTION" column
medications_ct1.csv	Cross Tabulation operation on "DESCRIPTION" and "REASONDESCRIPTION"
medications_nuvc1.csv	Unique Values Count on "DESCRIPTION" on dataframe filtered by NaN values in "REASONDESCRIPTION"
medications_unwanted.csv	Count of Unwanted data in the dataframe

Table 7: Files with results of EDA on medications.csv

- **Hydrochlorothiazide 25 MG Oral Tablet** is the most common medication (see *medication_uvc1.csv*).
- Most medications are prescribed for minor medical reasons i.e. "REASONDESCRIPTION" having a **NaN value**, the second most common reason is **Hypertension** (see *medication_uvc2.csv*).
- One medication could be used to treat different conditions (see *medications_ct1.csv*)
- In the absence on of any major medical reason i.e. "REASONDESCRIPTION" column has **NaN value** the most frequent entry in the data base is **Nitroglycerin 0.4 MG/ACTUAT Mucosal Spray** (see *medication_nuvc1.csv*)
- It was observed that "REASONCODE" contains some **NaN values**, some values in the "PAYER COVERAGE" and "TOTAL COST" columns are also **0**, although it is possible for a patient to have no health coverage, the possibility of "TOTAL COST" column having **0** might point towards the presence of bad data (see *medications_unwanted.csv*). thus a further investigation is required to determine if ther is any bad data in the *medication.csv*.

2.1.8 observations.csv

File	Operations
observations_uvc1.csv	Unique Values Count on "DESCRIPTION" column
observations_unwanted.csv	Count of Unwanted Data in the Dataframe

Table 8: Files with results of EDA on observations.csv

- **Pain severity - 0-10 verbal numeric rating [Score] - Reported** is the most commonly performed procedure.
- Based on the nature of the data it is difficult to draw any meaningful conclusion, thus a further analysis is required to draw meaningful insights.
- Based on an initial analysis it was observed that "ENCOUNTER" and "UNITS" column has some **NaN values** and some values in the "VALUE" column are also zero. Thus based on an initial analysis it can be stated that *observations.csv* does have some unwanted data.

2.1.9 procedures.csv

File	Operation
procedures_uvc1.csv	Unique Values Count on "DESCRIPTION" column
procedures_uvc2.csv	Unique Values Count on "REASONDESCRIPTION" column
procedures_ct1.csv	Cross Tabulation operation on "DESCRIPTION" and "REASONDESCRIPTION"
procedures_nuvc11.csv	Unique Values Count on "DESCRIPTION" on dataframe filtered by NaN Values in "REASONDESCRIPTION"
procedures_unwanted.csv	Count of unwanted data on dataframe

Table 9: Files with results of EDA on procedures.csv

- **Medication Reconciliation (procedure)** is the most frequently performed procedure (see *procedures_uvc1.csv*).
- Most of the procedures are performed due to **Normal Pregnenecy** (see *procedures_uvc2.csv*)
- It was observed that "REASONDESCRIPTION" does contain some **NaN Values** (see *procedures_nuvc1.csv*). However these cannot be regarded as missing values as this represents the cases where there is no major medical reason.
- Even in the absence of any major medical reason it was observed that **Medication Reconciliation (procedure)** was most frequently performed procedure(see *procedures_nuvc1.csv*) however it must be observed that in the absence of any major medical reason i.e. "REASONDESCRIPTION" column has a **NaN Value**.
- On a closer obervation it could be observed that a procedure could be performed for multiple medical reasons (see *procedures_ct1.csv*)