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

Friday 26th February, 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

After Performing EDA on *allergies.csv* the following was discovered:

- A unique value count was performed on "DESCRIPTION" column to determine the most common and least common type of allergy, the results of this operation are recorded in *allergy_DESCRIPTION.csv*. Based on the observation of *allergy_DESCRIPTION.csv* it was observed that allergy to mould is the most common tyoe of allergy whereas allergy to soya is the least common type of allergy.

- Unique value count was also performed on "PATIENTS" column and the results are recorded in *allergies_patient.csv*. Based on the enteries in *allergies_PATIENT.csv* it was observed that it is possible for a patient to have multiple allergies, however it contains information about allergies in the past and ongoing allergies as well. Thus a further, investigation is required to determine if a patient could have multiple ongoing allergies.

- To investigate if it is possible for a person to have multiple ongoing allergies the dataframe in *allergies.csv* was filtered by null values in the "STOP" column and then a unique value caount was performed on the "PATIENT" column. The results of this operation were stored in *allergies_ongoing.csv*. After examing the enteries in *allergies_ongoing.csv* it can be observed that it is possible for a patient to have multiple ongoing allergies.

- Record of unwanted data in the *allergies.csv* could be found in the file named *allergies_unwanted.csv*. There are a few NaN values in the stop column of *allergies.csv*, this could be attributed to the ongoing allergies of a patient and are not an anamoly

2.1.2 careplans.csv

After performing EDA operations on ***careplans.csv*** the following was discovered:

- A unique value count was performed on "PATIENT" column of *careplans.csv* and the results were stored in *careplans_PATIENT.csv*. After observing the enteries in *careplans_PATIENT.csv* it was observed the it is possible for a patient to have multiple careplans, however further investigation is required to check if it is possible for a patient to have multiple ongoing careplans.

- A unique value count was performed on "PATIENT" column of *careplans.csv* after filtering it by null values in the "STOP" column. The results of this operation are stored in *careplans_ongoing.csv*, based on the data in the afromentioned file it was observed that it is possible to have single patient id listed against multiple ongoing careplans

- 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.

- The results of performing a unique value count on "DESCRIPTION" column are stored in *careplans_DESCRIPTIONS.csv*. Based on these results it can be observed that careplans are most commonly used for respiratory therapy.

- Similarly a unique value count was also performed on "REASONDESCRIPTION" column resulting in the formation of *careplans_REASONDESCRIPTION.csv*. After studying the data in *careplans_REASONDESCRIPTION.csv* it was observed that most frequently occuring entry in the "REASONDESCRIPTION" was Acute Bronchitis(Disorder) which is a respiratory disorder thus justifying the fact that Respiratory Therepy is the most commonly occuring value in the "DESCRIPTION" column.

- It was observed that *careplans.csv* contain some entries where the "REASONDESCRIPTION" was empty but "DESCRIPTION" did contained some entries, on closer observation it was observed that whenever the "REASONDESCRIPTION" column was empty the entries in the "DESCRIPTION" column did not reflected any major condition, thus it was decided to observe how the patients are using the careplans when there was not a major medical illness, to perform this operation the dataset was filtered by null values in the "REASONDESCRIPTION" column and unique value count was perfomed on the "DESCRIPTION" column the resultant data was stored in *careplans_REASONDESCRIPTION_null.csv*. After performing this operation it was observed that when there was no primary medical issue the careplans are most commonly used for Self Care Interventions.

- The file named *careplans.csv* was also analyzed for unwanted data. As per the entries in *careplans_unwanted.csv* the only unwanted data present in*careplans.csv* were the NaN values present in the *"STOP","REASONCODE"* and *"REASONDESCRIPTION"* column. This cannot be treated as null values. The empty columns in the *"STOP"* column represent that the care plan is still ongoing whereas the NaN values in the *"REASONCODE"* and *"REASONDESCRIPTION"* column represnt the cases when careplans are used for no major medical reason.

2.1.3 conditions.csv

Based on the enteries in *conditions.csv*, the following eda approach was followed:

- A unique value count was performed on the "DESCRIPTION" column to discover the most and least common conditions. As per the entries in *conditions_DESCRIPTION.csv* it was discoverd that Viral Sinusitis (disorder) was the most common condition

- 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_STOP_null.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_STOP_null.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_psudo_chronic.csv*. The resultant conditions were recorded in *conditions_chronic.csv*

2.1.4 encounters

encounter.csv has multiple relationships with other files hence it is important that a more detailed analysis is performed in this file. Based on the entries in this file the following eda techniques were employed:

- Based on a unique value count of the "REASONDESCRIPTION" column, it was observed that most frequent entry in the "REASONDESCRIPTION" column is a null value and the second most frequent value is Normal Pregenancy. The results of unique value count are stored in careplans_REASONDESCRIPTION.csv

- Unique value count was also performed on "DESCRIPTION" column. It was observed that the most common entry in the "DESCRIPTION" column is well child visit(procedure). The results of this operation are stored in *encounters_DESCRIPTION.csv*.

- After performing a unique value count on the "ENCOUNTERCLASS" it was discovered that ambulatory is the most common value i.e. for most of he medical encounters the patient walked in for the procedure, assesment, consultation etc.

- *encounters.csv* was also analyzed for unwanted data and the output was stored in *encounters_unwanted.csv*. It was observed that only "REASONCODE" and "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 t "REASONCODE" and "REASONDESCRIPTION" column need to be investigated further.

- To investigate the entries where "REASONDESCRIPTION" is NaN value it was decided to filter the database by NaN values in the "REASDESCRIPTION" column and perform a unique values count on the "DESCRIPTION" and "ENCOUNTERCLASS" column, the results of these operations are stored in *encounters_RESAONDESCRIPTION_null_DESCRIPTION.csv* and *encounters_REASONDESCRIPTION_null_ENCOUNTERCLASS.csv*.

- While comparing the entries in *encounters_REASONDESCRIPTION_null_DESCRIPTION.csv* and *encounters_DESCRIPTION.csv* it can be noted that some values have the same count in both the files thus indicating that all the enteries with that particular value occur when the "REASONDESCRIPTION" column has NaN value. It can also be noted *encounters_REASONDESCRIPTION_null_DESCRIPTION.csv* has 33 unique values whereas *encounters_DESCRIPTION.csv* has 54 unique entries, thus based on this observation we can conclude that some values only occur when the "REASONDESCRIPTION" column has some value i.e. when the REASONDESCRIPTION column is not NaN.

- A similar observation can be made for *encounters_REASONDESCRIPTION_null_ENCOUNTERCLASS.csv* and *encounters_ENCOUNTERCLASS.csv*. Although the number of unique values in both the files are

same the frequency count of these values are different thus we can say that even when the "REASONDESCRIPTION" column has NaN values all type of values in the "ENCOUNTERCLASS" column can be observed.

- A further analysis is required on enteries where the "REASONDESCRIPTION" column has NaN values.
- The "ENCOUTERCLASS" column has several unique values, we are particularly interested in the encounters where "ENCOUTERCLASS" is emergency and urgentcare.
- Two cross tabulation operations were performed between three columns and the results were stored as csv. The first cross tabulation was performed between "DESCRIPTION" and "REASONDESCRIPTION" column and the results were stored in *encounters_crosstab_REASONDESCRIPTION_ENCOUNTERCLASS.csv*. Another cross tabulation operation was performed between "DESCRIPTION" and "ENCOUNTERCLASS" and the results were stored in *encounters_crosstab_REASONDESCRIPTION_ENCOUNTERCLASS.csv*. Based on the results of cross tabulation data it can be clearly seen that several values of the "DESCRIPTION" and "REASONDESCRIPTION" column are distributed across various types of "ENCOUNTERCLASS" values. A further investigation is required for cases which are not purley of type **emergency** and **ungertcare**

2.1.5 imaging_studies.csv

Based on the entries of *imaging_studies.csv* the following eda approach was followed:

- A unique value count was performed on "BODYSITE_DESCRIPTION" column and the results were stored in *imaging_studies_BODYSITE_DESCRIPTION.csv*. It was discoverd that Thoracic Structutre (Body Study) is the most examined body structure.
- Unique value count was also performed on "SOP_DESCRIPTION" column, based on the data and the results were stored in *imaging_studies_SOP_DESCRIPTION.csv*. Based on the enteries of this file it was discovered that Digital X-rays are the most commonly performed procedures.
- A cross tabulation operation was performed between "BODYSITE_DESCRIPTION" and "SOP_DESCRIPTIO and the results were stored in *imaging_studies_crasstab_BODYSITE_DESCRIPTION_SOP_DESCRIPTION.cs* Based on the data in this file it can be concluded that only one type of tests are being done on the body parts except for Thoracic Structure. It must also be noted that that this file also contains an entry named ***thoracic*** which could be similar to ***Thoracic Stucture(Body Structre)***
- Another cross tabulation operation was performed between "MODALITY_DESCRIPTION" and "SOP_DESCI and the results were stored in *imaging_studies_crosstab_MODALITY_DESCRIPTION_SOP_DESCRIPTION.c* However the results of this cross tabulation was not of any significant importance. It must be noted that "MODALITY_DECSRIPTION" column only contains the category under which the imaging study falls.
- *imaging_studies.csv* was also analyzed for unwanted data and the results were stored in *imaging_studies_unwant* Based on the entries of *imaging_studies_unwanted.csv* it can be said that *imaging_studies.csv* does not contain any unwanted data.

2.1.6 immunizations.csv

Based on the data in *immunization.csv* the following eda approach was adopted:

- A unique value count was performed on the "DESCRIPTION" column and the results were stored in *immunization_DESCRIPTION.csv*. Based on the description it was discoverd that ***Influenza seasonal injectable preservative free*** was the most frequently occurring entry.
- *immunization.csv* was also analyzed for unwanted data. No unwanted data was found in *immunization.csv*.

2.1.7 medications.csv

Based on the data in *medications.csv* the following eda approach was adopted:

- A unique value count was performed on "DESCRIPTION" column and the results were stored in *medications_DESCRIPTION.csv*. Based on the data present in *medications_DESCRIPTION.csv*, it was discovered that ***Hydrochlorothiazide 25 MG Oral Tablet*** was the most common entry in the file.
- A unique value count was also performed on the "REASONDESCRIPTION" column and the results were stored in *medications_REASONDESCRIPTION.csv*. Based on the data it can be stated that **null** is the most commonly occuring value and **Hypertension** is the second most commonly occuring value.
- To determine what medication is used for what purpose a cross tabulation operation was performed on "DESCRIPTION" and "REASONDECSRIPTION" column and the results were stored in *medications_crosstab_DESCRIPTION_REASONDESCRIPTION.csv*. Based on the data in this it was discovered that one medication could be used to treat different conditions.
- It was observed that a lot of entries in "REASONDESCRIPTION" column are null values hence the dataframe was filtered by null values and unique values count was performed on the "DESCRIPTION" column, the results of this operation are stored in *medications_REASONDESCRIPTION_null_DESCRIPTION* The most frequent entry in in the DESCRIPTION column when REASONDESCRIPTION is null is **Nitroglycerin 0.4 MG/ACTUAT Mucosal Spray.**

2.1.8 observations.csv

Although it was difficult to draw any conclusions from *observations.csv*, but the following approach was adopted to perform EDA :

- A unique values count was performed on the "DESCRIPTION" column and the results were stored in *procedures_DESCRIPTION.csv*. Based on the entries in the file it can be observed that **Pain severity - 0-10 verbal numeric rating [Score] - Reported** is the most frequently occurring value.
- A further analysis needs to be performed to extract meaningful features from this dataframe.

2.1.9 procedures.csv

Based on the entries in *procedures.csv* the following EDA approach was followed:

- A unique values count was performed on "DESCRIPTION" column, the results were stored in *procedures_DESCRIPTION.csv*. Based on this operation we can conclude that **Medication Reconciliation (procedure)** is the most frequent value
- A unique value count was also performed on "REASONDESCRIPTION" column and the results were stored in *procedures_REASONDESCRIPTION.csv*. Based on the results of this operation it can be **Normal Pregnenecy** is the most occuring value.
- To understand which procedure are performed for which reason a cross tabulation operation was performed between "DESCRIPTION" and "REASONDESCRIPTION" column and the results were stored in *procedures_crosstab_DESCRIPTION_REASONDESCRIPTION.csv*. Based on the results it was observed that one procedure could be performed for several different conditions.