**Suspect Analysis -**

**HCC Recommendation System**

**User Manual**

Contents

[Section 1: Project Overview 3](#_Toc134718802)

[Section 2: Data Preparation 4](#_Toc134718803)

[Section 3: Modelling – Surprise Library – (Optional step) 6](#_Toc134718804)

[Section 4: Modelling – cosine similarity- memory based 7](#_Toc134718805)

[Section 5: Modelling – cosine similarity – model based 8](#_Toc134718806)

[Section 6: RAF Calculation 9](#_Toc134718807)

# Section 1: Project Overview

Healthcare costs in the United States continue to rise year after year, making it imperative for healthcare payers to find ways to reduce the financial burden on both the government and the beneficiaries. Hierarchical Condition Category (HCC) is a system used by healthcare payers to identify and evaluate the health status of beneficiaries. The HCC system is a risk-adjustment methodology that helps payers to assess the expected costs of providing care to beneficiaries based on their medical conditions.

The HCC system works by assigning a score to each beneficiary based on their diagnosis codes, which are typically captured in medical claims data. Diagnosis codes are standardized codes that represent a beneficiary's medical condition. These diagnosis codes then mapped to HCC codes. Based on these HCC codes and other demographic details (eg. – Age, gender, health benefits, disability condition etc.) the risk score is calculated. Average risk score on each demographic area is then used to find the policy premium for that demographic.

The HCC categories are hierarchical, which means that more severe conditions are assigned to higher categories, and less severe conditions are assigned to lower categories.

The HCC score for each beneficiary is calculated by summing the scores assigned to the categories they belong to. The higher the HCC score, the more severe the beneficiary's health conditions, and the more costly it is expected to provide care for them.

The purpose of this project is to develop a recommendation engine that can suggest HCC codes to beneficiaries. The engine will help to reduce undercoding, which occurs when healthcare providers do not capture all of the beneficiary's medical conditions accurately. Undercoding can lead to an inaccurate assessment of a beneficiary's health status, which can result in a lower HCC score and lower payments to healthcare payers. By accurately identifying a beneficiary's health status and recommending appropriate HCC categories, the engine will help to increase the HCC score, which will ultimately increase the payer's profit.

The engine will work by using machine learning algorithms to analyse the beneficiary's medical claims and demographic’s data. It will then suggest appropriate HCC categories based on these details.

In conclusion, this project aims to develop a recommendation engine that will improve the accuracy of HCC scores and reduce undercoding, ultimately leading to increased payments to healthcare payers. By leveraging machine learning algorithms and medical claims data, the engine will provide accurate and timely recommendations for HCC categories, helping to ensure that beneficiaries receive appropriate care and payers receive accurate reimbursements.

# Section 2: Data Preparation

All the python codes of this section is present at General - Suspect analysis - Risk adjustment\Health Care HCC recommendation\codes\ Step 1 Data Prepration.ipynb

It is a necessary step

**2.1 Raw File Details**

In this project these Raw files are used to create master file –

|  |  |
| --- | --- |
| **Data Name** | **Description** |
| **1.     Test\_Beneficiarydata-1542969243754.csv** | Beneficiary level data |
| **2.     Test\_Inpatientdata-1542969243754.csv** | Inpatient Claims data |
| **3.     Test\_Outpatientdata-1542969243754.csv** | Outpatient Claims data |
| **4.     Train\_Beneficiarydata-1542865627584.csv** | Beneficiary level data |
| **5.     Train\_Inpatientdata-1542865627584.csv** | Inpatient Claims data |
| **6.     Train\_Outpatientdata-1542865627584.csv** | Outpatient Claims data |
| **7.     icd9\_to\_icd10\_cm\_mapping.csv** | ICD9 to ICD10 mapping file |
| **8.     PY 2024 Proposed Clinical Revision Part C Model ICD-10 Mappings.xlsx** | ICD10 to HCC mapping file |
| **9.     SDOH\_working\_file.xlsx** | Social determinant of health data at zip code level |
| **10.  HCC\_Factors\_unpivoted.xlsx** | risk adjustment factors for each disease |

**2.2 Claim’s data merging**

\*\*append

df\_in\_train

df\_in

df\_in\_test

df\_patients

df\_out\_train

df\_out

df\_out\_test

Merge

On – Beneficiary ID

df\_ben\_train

df\_ben\_test

df\_ben

df\_patients\_ben

**2.3 ICD9 -> ICD10 -> HCC mapping**

df\_patients\_ben has ICD 9 codes. Firstly ICD 9 codes were converted into ICD 10 codes using df\_icd9\_icd10 then ICD 10 codes mapped into HCC codes using df\_icd10\_hcc.

**2.4 Data Cleaning and Feature Engineering**

Dropped some unnecessary columns and created “Age” and “claim duration” columns.

**2.5 Data mockup**

Since the major purpose of this project is to reduce the undercoading problem and improve the risk adjustment factor. So, we have mocked up the data and introduced those columns which are necessary to calculate risk score.

1. community\_institutional – (Community, institutional)
2. Disability\_condn – (Non\_disabled, Non\_originally\_disabled, Originally\_disabled)
3. Benefits – (NonDual, FBDual, PBDual)
4. medicaid\_flag – (1,0)

\*\* while mocking the data standard ratios has taken into consideration so that I will reflect the real-world conditions.

Based on the HCC codes and using standard coding of chronical disease. Disease interactions and disease disability interaction columns has been created.

**2.5 SDoH Data merging**

Social determinant of health (SDoH) data has also been merged in df\_patients\_ben\_final. Since df\_patients\_ben\_final data was on state level and SDoH data was on zip code level so based on population density of each zip code in each state this SDoH data has been merged to df\_patients\_ben\_final.

**2.5 Data cleaning and exporting**

One beneficiaries’ all information is same but it can have multiple HCC’s so to breakdown the data at beneficiary HCC level and to create a differentiation between them another column ‘other\_disease\_mean\_hcc\_factor’ is created which is a mean factor of all other HCC scores.

Useful columns are selected and rearranged. Dropped those rows which has empty HCC’s. Then the data is exported on the name of –

**‘master\_data\_for\_modelling.csv’**

Full dataset is exported to create a PowerBI Dashboard on the name of –

**‘data\_for\_powerBI.csv’**

# Section 3: Modelling – Surprise Library – (Optional step)

All the python codes of this section is present at General - Suspect analysis - Risk adjustment\Health Care HCC recommendation\codes\ Step 2 modelling surprise SVD.ipynb

It is NOT a necessary step. it was done initially but since the results were not good so, the cosine similarity approach is then used.

STEPS: –

1. importing library and datasets
2. converting categorical columns into strings
3. scaling and one hot encoding of data
4. missing values imputation with median
5. finding top features based on decision tree
6. creating a classification model on top features to see its performance (extra step)
7. Apply different techniques (PCA, LDA, MDA, FAMD) to combine the variables into a single factor since Surprise model takes only one factor. Finally creating a data frame that contains 3 columns ('BeneID','HCC','factors')
8. Load the data into Surprise's Dataset object
9. Train the algorithm on the data
10. Predicting the recommendations and checking the results

Since the model is not performing well so will not use this model for recommendations

Possible reasons of inaccuracy-

1. we are using only 19 features to find the 1 master feature.

2. even this master feature is only able to capture 74% variability from 19 features whereas total features are 81

# Section 4: Modelling – cosine similarity- memory based

All the python codes of this section is present at General - Suspect analysis - Risk adjustment\Health Care HCC recommendation\codes\ Step 3 modelling cosine similarity memory based.ipynb

It is NOT a necessary step. it was done after surprise libraries’ approach but since it is an memory based approach, so it can not generate recommendations for new beneficiary that’s why in the next step this model is converted into memory based approach and some extra refinements also been done. To see that follow section 5.

Cosine similarity technique was used to find similar accounts based on all relevant features.

STEPS: -

1. importing library and datasets
2. converting categorical columns into strings
3. scaling and one hot encoding of data
4. missing values imputation with median
5. finding calculating user\_user cosine similarity matrix
6. looping through each beneficiary and finding all similar beneficiaries whose similairty socre is more than the threshold similarity (0.7 can be modified). Creating a data frame having Target\_beneficiary, Simialr\_beneficiary and similarity score
7. adding list of target beneficiaries HCC and similar beneficiaries HCC
8. finding common HCC and their count
9. second level of sorting is done on the basis of common HCC count for each target beneficiary.
10. Then individual recommendations has been calculated.

Individual recommendations = similar HCC – Target HCC

1. Item-Item similarity has also been calculated but that was not performing better so not taken into consideration

# Section 5: Modelling – cosine similarity – model based

All the python codes of this section is present at General - Suspect analysis - Risk adjustment\Health Care HCC recommendation\codes\ Step 4 modelling cosine similairty model based.ipynb

It is a necessary step. it can be run directly after data preprocessing step. in this step is done to convert the memory-based approach to model-based approach by using singular value decomposition. Also, some advance features are created and some risk adjustment score based columns has been merged on target beneficiary to calculate the risk score before and after recommendations. This is done in next step.

Cosine similarity technique was used to find similar accounts based on all relevant features.

STEPS: -

1. importing library and datasets
2. converting categorical columns into strings
3. scaling and one hot encoding of data
4. missing values imputation with median and concatenating BeneID to transformed data
5. creating data at beneficiary level by grouping it at beneficiary level and finding mean of all median of all values
6. defining a python class named CosineRecommendation in order to convert memory based to model-based approach
7. making an instance of this class then fit the model on input data.
8. Splitting the test data on 500 datasets row wise. Looping through each of the split dataset to apply predict method on these datasets that predicts the similar beneficiaries for each target beneficiary with similarity score more than 0.8 and returns a dataset contains Target\_beneficiary, Simialr\_beneficiary and similarity score
9. adding list of target beneficiaries HCC and similar beneficiaries HCC
10. finding common HCC and their count
11. second level of sorting is done based on common HCC count for each target beneficiary then top 5 similar beneficiaries has taken.
12. Then individual recommendations has been calculated.

Individual recommendations = similar HCC – Target HCC

1. Combined recommendations have been created for each target beneficiary. Sorting of these combined recommendations has been done by frequency count of recommended HCC. Whom so ever HCC was recommended by more and more similar beneficiaries comes on top.
2. Since there can be a large number of combined recommendations possible for a single target beneficiary so taking top 5 recommendations and putting them into combined\_recommedned\_HCC\_trimmed column.
3. merging beneficiary level data in recommendations data to calculate RAF score.

Result of this step will be exported inside data/processed\_data as an csv file named as –

‘all\_recommendations\_0.8\_top5.csv’

# Section 6: RAF Calculation

All the python codes of this section is present at General - Suspect analysis - Risk adjustment\Health Care HCC recommendation\codes\ Step 5 Risk Adjustment Factor.ipynb

It is a necessary step. it can be run after Step 4 modelling cosine similarity model based. this step is done to calculate the risk score before and after recommendations.

For the calculation of risk score before recommendation and after recommendations the steps taken are given as below. There is an assumption that after recommending HCC’s to all the beneficiaries top 2 HCC’s of every beneficiary is found positive. So for calculating risk score 2 extra HCC’s are taken into account along with the True HCC’s that a beneficiary already have.

STEPS: -

1. libraries and datasets are imported.
2. In data all\_recommendations\_0.8\_top5.csv (result of last step) and HCC\_Factors\_unpivoted.xlsx taken. One thing to note is this excel file has multiple sheets in it so all the sheets has been imported in different data frames and different data frames has been formed. These are given as
   1. RAF\_Age\_gender
   2. RAF\_Medicaid\_OrgnlDsbl
   3. RAF\_Disease\_Coefficients
   4. RAF\_Disease\_intraction
   5. RAF\_Disabled\_Disease\_Int
   6. RAF\_Payment\_HCC\_Counts
   7. hirarcial\_conditions
3. since if a pandas dataframe is exported into a csv file and it contains lists inside a column then it is stored as string and when we import in a notebook it will work as string. So these strings are converted into lists.
4. As mentioned above due to top 2 HCC assumption taking only top 2 HCC from combined\_recommedned\_HCC\_trimmed.
5. Different columns of different datasets are converted into required format so that those can be merged. These datasets are RAF\_Age\_gender, RAF\_Disabled\_Disease\_Int, RAF\_Disease\_intraction, RAF\_Disease\_Coefficients, RAF\_Payment\_HCC\_Counts, all\_recommendations
6. Since recommended HCC’s has been added in data that’s why disease interactions and disease disability interactions has been again calculated
7. Dropped HCC’s due to hierarchal conditions.
8. Made a column for Age and gender factor by merging all\_recommendations and RAF\_Age\_gender on multiple keys through left join
9. Made a column for RAF\_Medicaid\_OrgnlDsbl by merging all\_recommendations and RAF\_Medicaid\_OrgnlDsbl on multiple keys through left join
10. To find sum of HCC factors before and after both the time data has been exploded and merged with RAF\_Disease\_Coefficients and them grouped on Target\_Beneficiary and found the sum of HCC\_factors
11. Made a column for RAF\_Disease\_intraction by merging all\_recommendations and RAF\_Disease\_intraction on multiple keys through left join
12. Similarly disease disability interactions has also been calculated
13. Risk factor due to HCC counts before and after also calculated
14. By summing the required columns RAF\_before and RAF\_after has been calculated.
15. Required columns selection and rearrangement has been done

Result of this step will be exported inside data/processed\_data as an csv file named as –

‘all\_recommendations\_0.8\_top5\_withRAF.csv’

Ernst & Young LLP

EY | Assurance | Tax | Transactions | Advisory

About EY

EY is a global leader in assurance, tax, transaction and advisory services. The insights and quality services we deliver help build trust and confidence in the capital markets and in economies the world over. We develop outstanding leaders who team to deliver on our promises to all of our stakeholders. In so doing, we play a critical role in building a better working world for our people, for our clients and for   
our communities.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. For more information about our organization, please visit ey.com.

Ernst & Young LLP is one of the Indian client serving member firms of EYGM Limited. For more information about our organization, please visit www.ey.com/in.

Ernst & Young LLP is a Limited Liability Partnership, registered under the Limited Liability Partnership Act, 2008 in India, having its registered office at 22 Camac Street, 3rd Floor, Block C, Kolkata - 700016

© 2018 Ernst & Young LLP. Published in India.

All Rights Reserved.

ED None

This publication contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither Ernst & Young LLP nor any other member of the global Ernst & Young organization can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication. On any specific matter, reference should be made to the appropriate advisor.