Predicting Length of Stay using MediCare claims data

Anay Dutta
Zohaib Sheikh
Pruthvinath

Overview

- Data: 4 different datasets:
  - Beneficiary summary (2008 - 2010)
  - Inpatient Claims

- Beneficiary summary Sample:
  - Beneficiary Summary sample data contains 32 variables across all the 3 years
  - Beneficiary Summary sample data has approx. 115k records each year

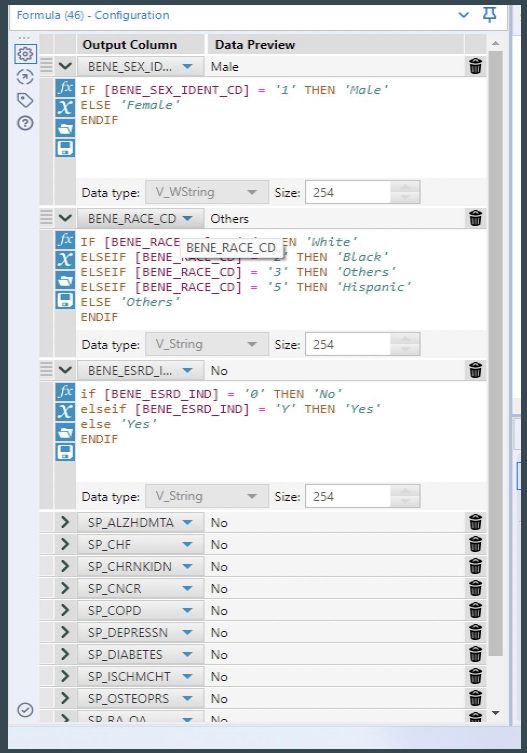
- Inpatient claims Sample
  - Inpatient Claims sample data contains 81 variables
  - Inpatient Claims sample data has approx. 66k records

Project objective:

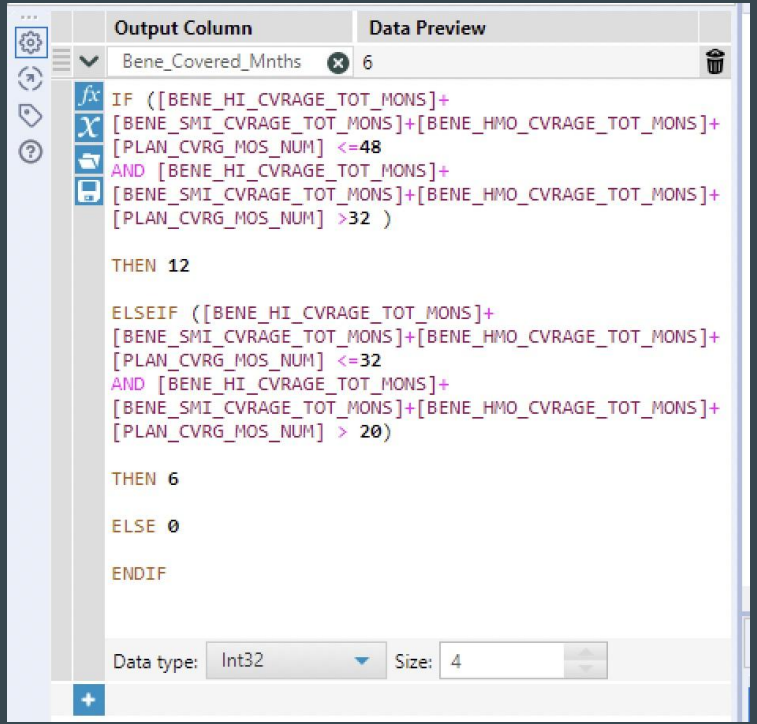
Predicting Length of stay based on Beneficiary summary and Inpatients claims data

Feature Engineering - Beneficiary Summary Data

- Appending 3 years data sets of Beneficiary summary (2008 - 2010)
- Converting the data type of required features.
- Converting different columns from numerical categories to understandable categories in English (this also helps in not misunderstanding the numerical categories as numerics in modelling)



- Combined different coverage months of part A, B, D into new single column i.e. Bene_Covered_Mnths into 3 categories - 0, 6 and 12 months.
- Then removed the above raw columns.

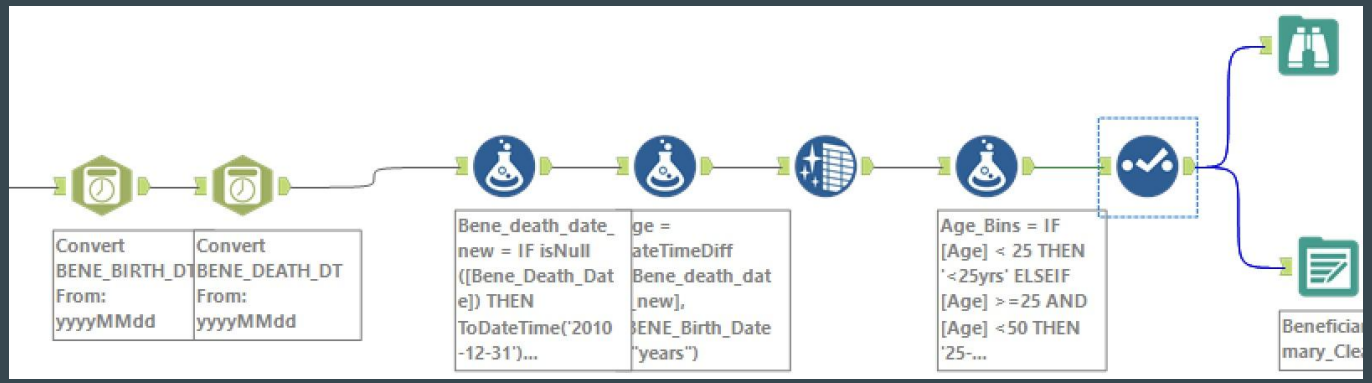


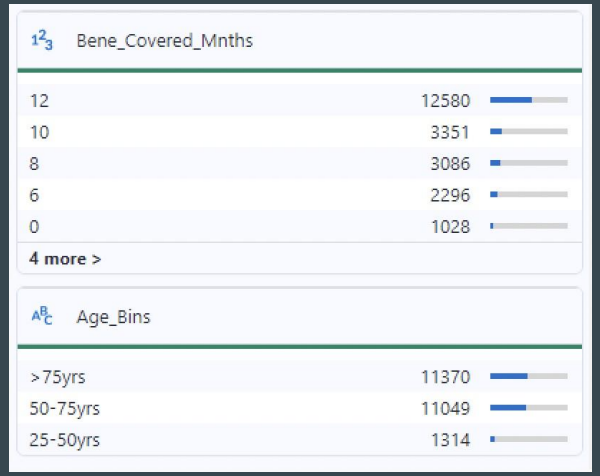
- Then summarize the whole 3 years data on ID as following:
  - Demographics like birth date, state, country, race etc. are same for all 3 years => took first occurrence.
  - Existing disease conditions are different for different years => took the mode of each condition for each individual.
  - Beneficiary covered months as an average.

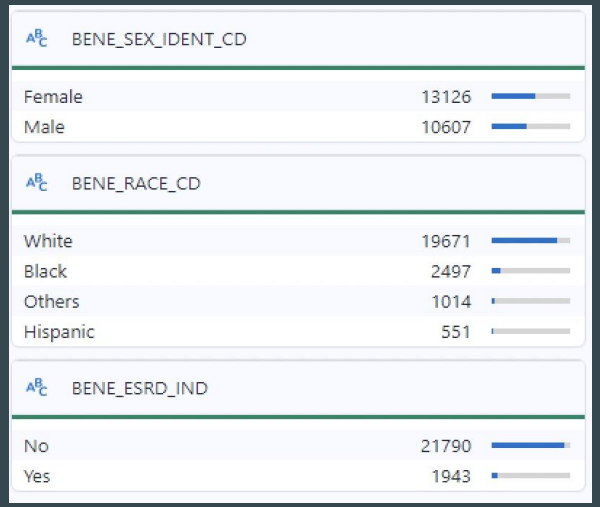
Actions: Add

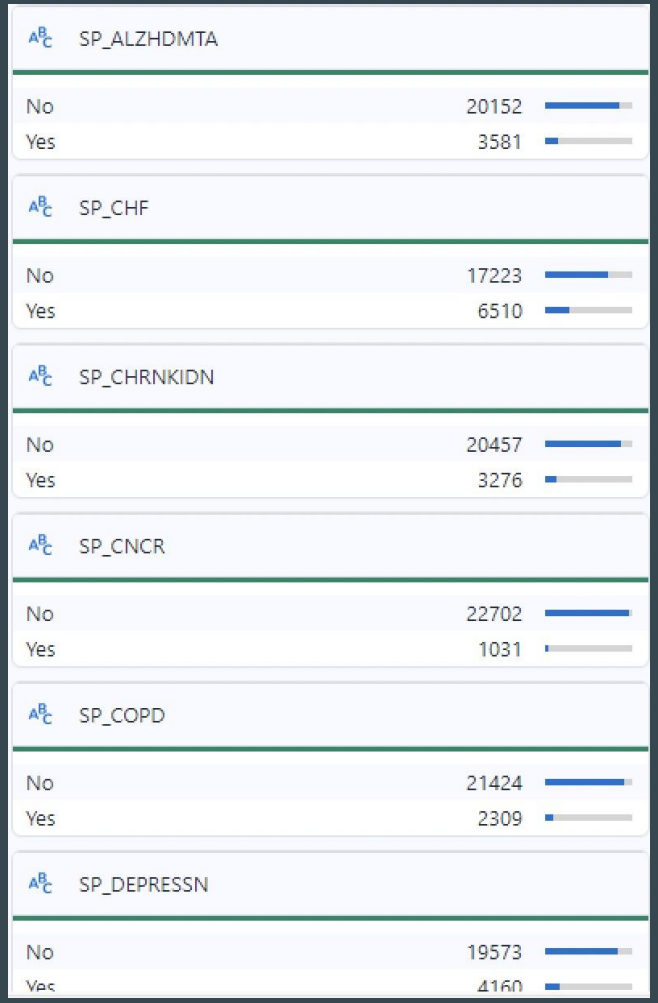
	Field	Action	Output Field Name
▶	DESYNPUF_ID	Group By	DESYNP...
	BENE_BIRTH_DT	First	BENE_BI...
	BENE_DEATH_DT	First	BENE_D...
	BENE_SEX_IDENT_CD	First	BENE_S...
	BENE_RACE_CD	First	BENE_R...
	BENE_ESRD_IND	First	BENE_E...
	SP_STATE_CODE	First	SP_STAT...
	BENE_COUNTY_CD	First	BENE_C...
	SP_ALZHDMTA	Mode	SP_ALZ...
	SP_CHF	Mode	SP_CHF
	SP_CHRNKIDN	Mode	SP_CHR...
	SP_CNCR	Mode	SP_CNCR
	SP_COPD	Mode	SP_COPD
	SP_DEPRESSN	Mode	SP_DEP...
	SP_DIABETES	Mode	SP_DIAB...
	SP_ISCHMCHT	Mode	SP_ISCH...
	SP_OSTEOPRS	Mode	SP_OST...
	SP_RA_OA	Mode	SP_RA_...
	SP_STRKETIA	Mode	SP_STR...
	Bene_Covered_Mnths	Average	Bene_C...

- Manipulating the birth and death dates (or with 2010-12-31) to find the Age and then converting age into age categories as <25 years, 25-50 years, 50-75 years and >75 years.
- Then removing null values in entire data set and also the reimbursement amounts columns as those amounts comes after the patient discharge => no role in estimating length of the stay.
- Atlast, output the beneficiary summary data set into a CSV file, which then be used in tests and modelling.



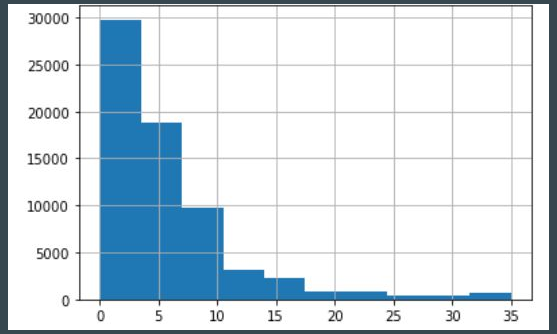

Bene_Covered_Mnths
12 12580
10 3351
8 3086
6 2296
0 1028
4 more >
Age_Bins
>75yrs 11370
50-75yrs 11049
25-50yrs 1314


BENE_SEX_IDENT_CD
Female 13126
Male 10607
BENE_RACE_CD
White 19671
Black 2497
Others 1014
Hispanic 551
BENE_ESRD_IND
No 21790
Yes 1943


SP_ALZHDMTA
No 20152
Yes 3581
SP_CHF
No 17223
Yes 6510
SP_CHRNKIDN
No 20457
Yes 3276
SP_CNCR
No 22702
Yes 1031
SP_COPD
No 21424
Yes 2309
SP_DEPRESSN
No 19573
Yes 4160

Feature Engineering- Inpatient Dataset

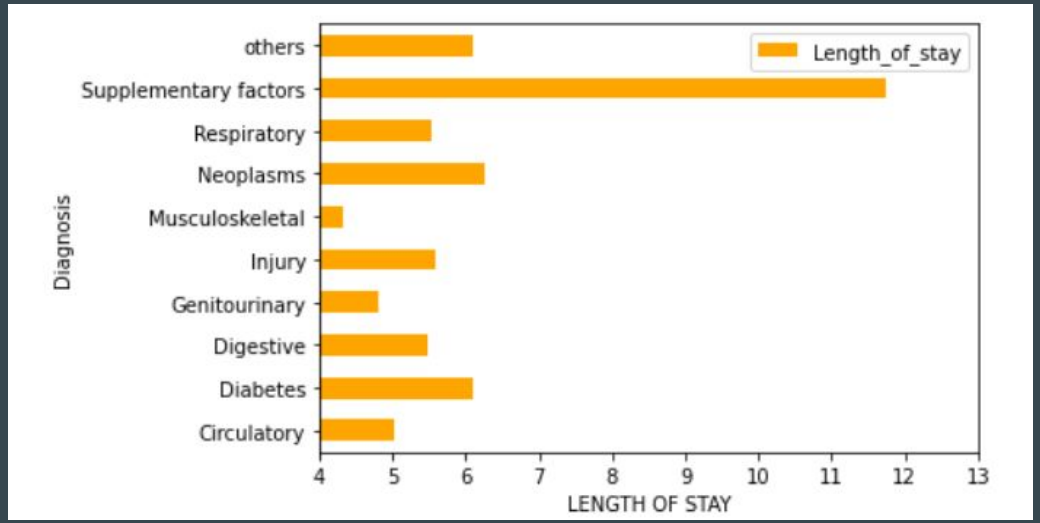
- Created a new column “Length of stay” using "CLM_FROM_DT", "CLM_THRU_DT"
- Removed the column “Segment” because of disproportionate values.
- Engineered “ADMTNG_ICD9_DGNS_CD” into 9 categories: Same was done for “ICD9_DGNS_CD_1-9”
- Removed all the columns with null values
- Columns taken in to account: DESYNPUF_ID, ADMTNG_ICD9_DGNS_CD, ICD9_DGNS_CD_1-9, Length_of_stay

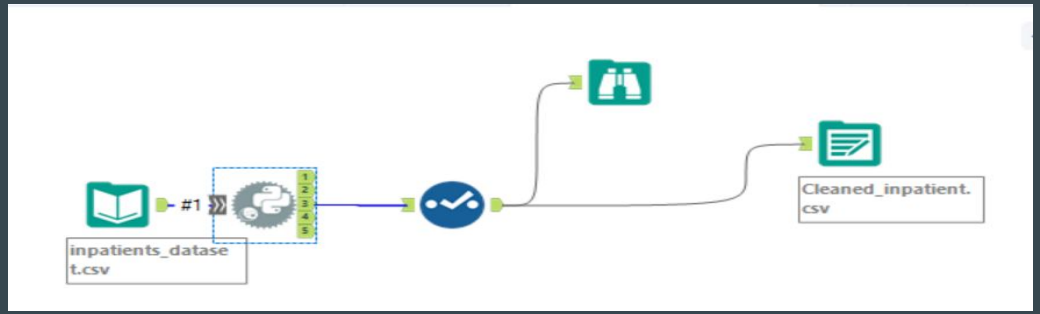



```
others                  19336
Respiratory             13416
Circulatory             11029
Digestive                5788
Musculoskeletal          5607
Injury                   4311
Genitourinary            3202
Supplementary factors    1937
Neoplasms                1589
Diabetes                  490
Name: ADMTNG_ICD9_DGNS_CD, dtype: int64
```

Data Exploration For Inpatient Dataset

Found that out of all the factors, Supplementary factors contribute towards the highest length of stay.

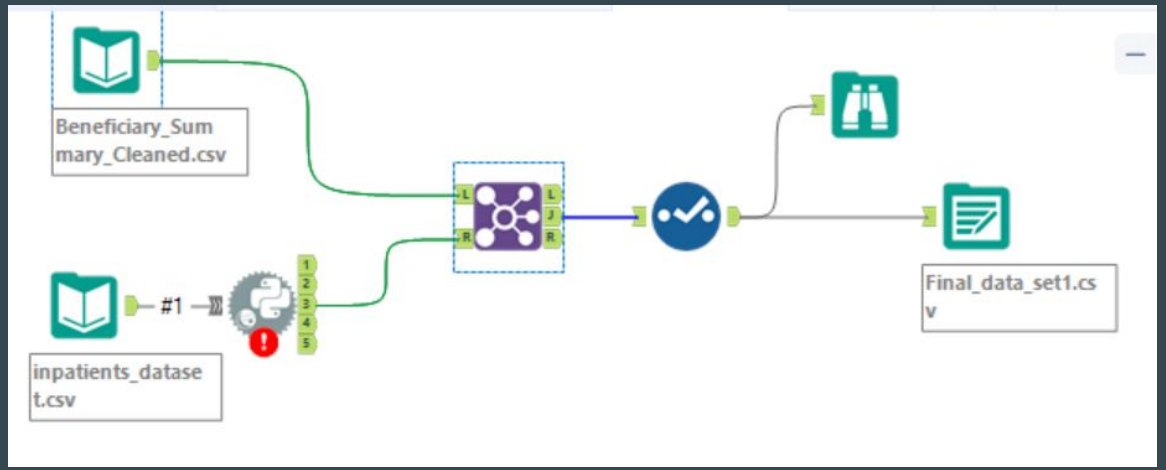




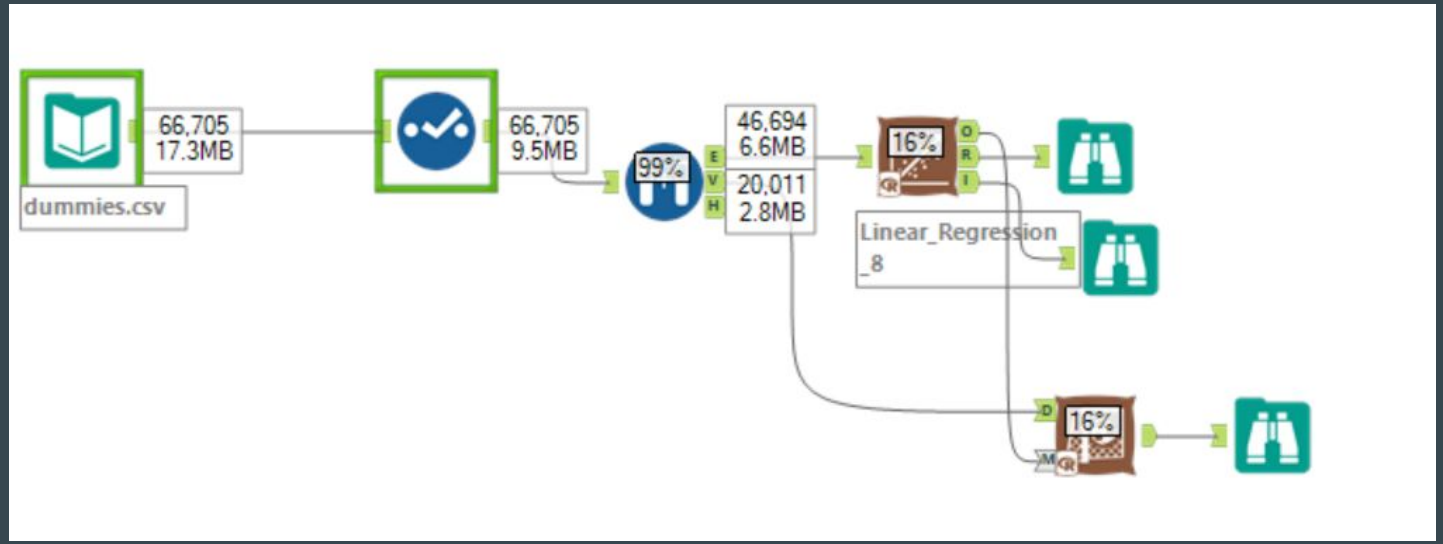
Final Dataset and model Building

1. We then inner joined our two datasets, which are Beneficiary dataset and Inpatient claims dataset to obtain.

2. In the next steps this final Data set was used to get a train dataset and a test dataset.



Modelling



We proceeded with a Linear regression model keeping the “Length of stay Numerical”

Adjusted R square of 0.725

Testing and Validation

Results

THANK YOU!!