

APAC – Data Science Use Case 1

Contents

Introduction	2
Performance-Grid	2
Healthcare Providers Data Science Project	3
Preamble.....	3
Problem Statement & Objectives.....	4
Dataset and Data Dictionary.....	5
How to download the data files	8
Section 1 – Problem Statements and Hypothesis	9
Section 2 – Data Engineering.....	10
Section 3 – Feature Engineering & Insights.....	11
Section 4 – Predictive Modelling & Evaluation.....	13
Section 5 – Model Operationalization	14
References	15

Introduction


APAC Data Science Performance -Grid has been designed for candidates being appeared in data science boot camp and long term program. This grid will have multiple projects by different healthcare market with various level of difficulty i.e. HPR market will have hospital readmission prediction project with several section and ranging difficulty from Level 1 to Level 6. It is expected from each candidate to score minimum by each market and by each section during program.

Each project is evaluated by Domain Experts and Data Scientists upon completion.

Performance-Grid

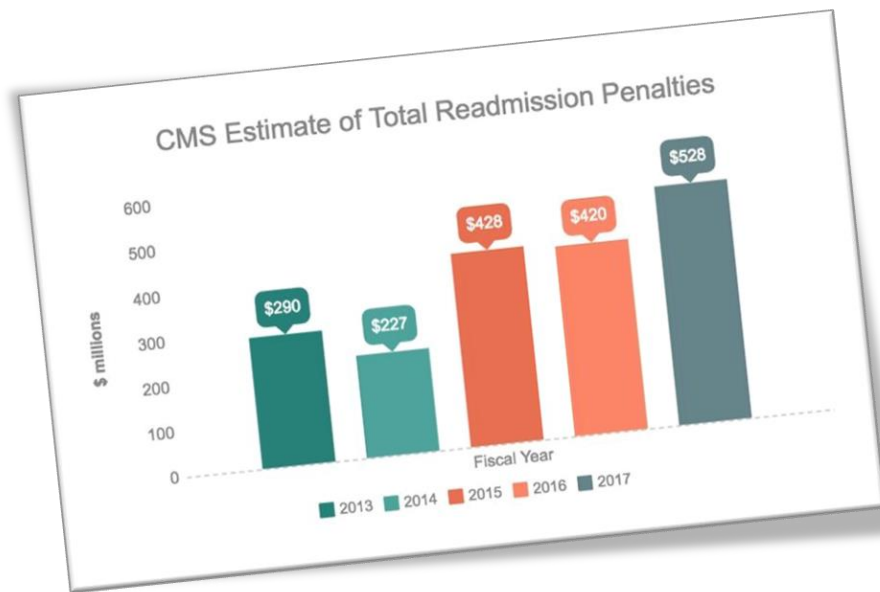
*L1-L6 are the questions to solve in each of grid and will fetch marks from 10-50 based on difficulty level.

Section /Market	HPR	HPP	HLS	HMT	Qualifying Score
<i>Problem Statements (hypothesis testing)</i>	L1 – 10 , L2 – 20 , L3 – 20 , L4 - 50				Sum (at least 2 problems in each market by each section)
<i>Data Engineering (data ingestion, data integration)</i>					
<i>Feature Engineering (Feature creation , insights,dimensionality reduction)</i>					
<i>Predictive Modelling (modelling ,evaluation, model bias, drift)</i>					
<i>Model Operationalization</i>					



Healthcare Providers Data Science Project

Project : Fewer Hospital U-turns



Preamble

Approximately 4.3 million hospital readmissions occur each year in the U.S., costing more than \$60 billion, with preventable adverse patient events creating additional clinical and financial burdens for both patients and healthcare systems.

Total Medicare penalties assessed on hospitals for readmissions increased to \$528 million in 2017, \$108 million more than in 2016. The increase is due mostly to more medical conditions being measured. Hospital fines will average less than 1 percent of their Medicare inpatient payments.





APAC candidates are expected to develop novel solutions to achieve the **Quadruple Aim of improving the patient experience of care, improving the health of**

populations, reducing cost, and improving clinical care provider satisfaction.

Problem Statement & Objectives

1. Use AI methodologies to predict unplanned hospital and Skilled Nursing Facilities admissions and adverse events within 30 days for Medicare beneficiaries, based on a data set of Medicare administrative claims data, including Medicare Part A (hospital) and Medicare Part B (professional services).
2. Develop innovative strategies and methodologies to: explain the AI-derived predictions to front-line clinicians and patients to aid in providing appropriate clinical resources to model participants

*There are four parts of Medicare: Part A, Part B, Part C, and Part D

-  Part A provides inpatient/hospital coverage
-  Part B provides outpatient/medical coverage
-  Part C offers an alternate way to receive your Medicare benefits
-  Part D provides prescription drug coverage

You can read more about these parts at [Medicare Coverage](#)

The Solution should prioritize explainable artificial intelligence solutions to help front-line clinicians understand and trust artificial intelligence-driven data feedback to target scarce resources and improve the quality of care.

It is expected from the model(s) that it should predict unplanned hospital and skilled nursing facility admissions within 30 days of discharge, **as well as adverse events like respiratory failure, postoperative pulmonary embolism, deep vein thrombosis, or sepsis.**

Dataset and Data Dictionary

The CMS linkable 2008–2010 Medicare Data Entrepreneurs’ Synthetic Public Use File (DE-SynPUF) was designed to create new type of file that would be useful for data entrepreneurs for software and application development and training purposes. The files preserve the detailed data structure and metadata of key variables at both the beneficiary and claim levels. However, the data are fully “synthetic,” meaning no beneficiary in the DE-SynPUF is an actual Medicare beneficiary. They are all synthetic beneficiaries meant to represent actual beneficiaries. In order to protect the privacy of beneficiaries and to greatly reduce the risk of re-identification, a significant amount of interdependence and co-variation among variables has been altered in the synthetic process. The synthetic process used significantly diminishes the analytic utility of the file to produce reliable inferences about the actual Medicare beneficiary population (i.e., univariate statistics and regression coefficients produced with the DE-SynPUF will be biased).

The CMS linkable 2008–2010 Medicare DE-SynPUF contains multiple files per year for multiple years. The DE-SynPUF contains multiple files per year for multiple years. The file contains synthesized data taken from a 5% random sample of Medicare beneficiaries in 2008 and their claims from 2008 to 2010. Each synthetic beneficiary was assigned a unique unidentifiable ID, DESYNPUF_ID, which is provided on each file to link synthetic claims to a synthetic beneficiary. This beneficiary ID carries no information about the enrollee or any patient records, and is provided solely for reference and data processing purposes.

For this synthetic sample of Medicare beneficiaries, the DE-SynPUF contains five types of files –

- **CMS Beneficiary Summary DE-SynPUF**
- **CMS Inpatient Claims DE-SynPUF**
- **CMS Outpatient Claims DE-SynPUF**
- **CMS Carrier Claims DE-SynPUF (also known as the Physician/Supplier Part B claims file)**
- **CMS Prescription Drug Events (PDE) DE-SynPUF**

Files of the same type contain the same sets of variables for each year. Variable names in the DE-SynPUF were kept the same as those in the actual Medicare data unless they were significantly coarsened to decrease re-identification risk.

In those cases, “SP_” was added to the original variable name for distinguishing. It is essential that the confidentiality of the Medicare beneficiaries are protected when producing Medicare public use files. The protection of such information makes it difficult to maintain the analytic utility of such an information rich data set. All variables in the DE-SynPUF are imputed/suppressed/coarsened as part of disclosure treatment.

The analytic utility of the data file differs based on the type and level of analysis being conducted:

- **Demographic:** The DE-SynPUF estimates of demographic characteristics (date of birth, date of death, sex, race, state, county) of the beneficiary population match the univariate frequency of the full population of beneficiaries enrolled in Medicare at any time during the 2008 year.
- **Clinical:** The DE-SynPUF estimates for clinical variables such as chronic conditions can provide researchers with bounds on how many cases with a specific condition are likely to be in the Medicare claims, which could be used to generate power calculations for a grant application.
- **Economic/financial:** The DE-SynPUF estimates for the economic and financial variables provide a lower bound for the true estimate of cost for the full population of beneficiaries enrolled in Medicare at any time during the 2008 year and costs for 2009 and 2010 for this 2008 beneficiary example.
- **Multivariate modeling:** The dynamic relationships between variables (demographic, health plan enrollment, clinical, economic/financial, and provider information) were altered, to limit reidentification risk. Therefore, analyses from multivariate modeling should be interpreted with caution. However, the programs and procedures employed in the multivariate modeling will function on the CMS Limited Data Sets or Identifiable Data prior to 2011.

Following sections describe variables in detail in the CMS Beneficiary Summary DE-SynPUF, the CMS Inpatient Claims DE-SynPUF, the CMS Outpatient Claims DE-SynPUF, the CMS Carrier Claims DE-SynPUF, and the CMS Prescription Drug Events (PDE) DE-SynPUF, respectively. A quick summary of the characteristics of DE-SynPUF is provided below.

Summary of Variables

The CMS Beneficiary Summary DE-SynPUF contains 32 variables. For variables available in denominator files, we kept the same variable name as in denominator files. Although variables in Beneficiary Summary DE-SynPUF were imputed and coarsened, for most variables, the format of the data values in the DE-SynPUF is the same as in the original data (e.g. the imputed county codes are valid county codes). In the few exceptions, “SP_” was added as prefix to the original variable name to distinguish those

Performance-grid

data items whose values no longer represent the typical values or the format of the original data field. Each record pertains to a synthetic Medicare beneficiary and includes:

#	Variable names	Labels
1	<i>DESYNPUF_ID</i>	DESYNPUF: Beneficiary Code
2	<i>BENE_BIRTH_DT</i>	DESYNPUF: Date of birth
3	<i>BENE_DEATH_DT</i>	DESYNPUF: Date of death
4	<i>BENE_SEX_IDENT_CD</i>	DESYNPUF: Sex
5	<i>BENE_RACE_CD</i>	DESYNPUF: Beneficiary Race Code
6	<i>BENE_ESRD_IND</i>	DESYNPUF: End stage renal disease Indicator
7	<i>SP_STATE_CODE</i>	DESYNPUF: State Code
8	<i>BENE_COUNTY_CD</i>	DESYNPUF: County Code
9	<i>BENE_HI_CVRAGE_TOT_MONS</i>	DESYNPUF: Total number of months of part A coverage for the beneficiary.
10	<i>BENE_SMI_CVRAGE_TOT_MONS</i>	DESYNPUF: Total number of months of part B coverage for the beneficiary.
11	<i>BENE_HMO_CVRAGE_TOT_MONS</i>	DESYNPUF: Total number of months of HMO coverage for the beneficiary.
12	<i>PLAN_CVRG_MOS_NUM</i>	DESYNPUF: Total number of months of part D plan coverage for the beneficiary.
13	<i>SP_ALZHDMTA</i>	DESYNPUF: Chronic Condition: Alzheimer or related disorders or senile
14	<i>SP_CHF</i>	DESYNPUF: Chronic Condition: Heart Failure
15	<i>SP_CHRNKIDN</i>	DESYNPUF: Chronic Condition: Chronic Kidney Disease
16	<i>SP_CNCR</i>	DESYNPUF: Chronic Condition: Cancer
17	<i>SP_COPD</i>	DESYNPUF: Chronic Condition: Chronic Obstructive Pulmonary Disease
18	<i>SP_DEPRESSN</i>	DESYNPUF: Chronic Condition: Depression
19	<i>SP_DIABETES</i>	DESYNPUF: Chronic Condition: Diabetes
20	<i>SP_ISCHMCHT</i>	DESYNPUF: Chronic Condition: Ischemic Heart Disease
21	<i>SP_OSTEOPRS</i>	DESYNPUF: Chronic Condition: Osteoporosis
22	<i>SP_RA_OA</i>	DESYNPUF: Chronic Condition: rheumatoid arthritis and osteoarthritis (RA/OA)
23	<i>SP_STRKETIA</i>	DESYNPUF: Chronic Condition: Stroke/transient Ischemic Attack
24	<i>MEDREIMB_IP</i>	DESYNPUF: Inpatient annual Medicare reimbursement amount
25	<i>BENRES_IP</i>	DESYNPUF: Inpatient annual beneficiary responsibility amount

Performance-grid

26	<i>PPPYMT_IP</i>	DESYNPUF: Inpatient annual primary payer reimbursement amount
27	<i>MEDREIMB_OP</i>	DESYNPUF: Outpatient Institutional annual Medicare reimbursement amount
28	<i>BENRES_OP</i>	DESYNPUF: Outpatient Institutional annual beneficiary responsibility amount
29	<i>PPPYMT_OP</i>	DESYNPUF: Outpatient Institutional annual primary payer reimbursement amount
30	<i>MEDREIMB_CAR</i>	DESYNPUF: Carrier annual Medicare reimbursement amount
31	<i>BENRES_CAR</i>	DESYNPUF: Carrier annual beneficiary responsibility amount
32	<i>PPPYMT_CAR</i>	DESYNPUF: Carrier annual primary payer reimbursement amount

Following user manual can be used for details about variables in files as well. It also lists out how to join different data elements by primary keys.



SynPUF_DUG_User_
Manual.pdf

How to download the data files

Reference : [Downloadable-Public-Use-Files](#)

Due to file size limitations, each data type in the CMS Linkable 2008-2010 Medicare DE-SynPUF is released in 20 separate samples (essentially each is a .25% sample). All claims for a particular beneficiary are in samples with the same number (i.e. all beneficiaries in sample 1 have all their claims in the sample 1 files). This design allows DE-SynPUF users who do not need the entire synthetic population of the DE-SynPUF to read in only as many samples as they desire.

A unique cryptographic identifier, DESYNPUF_ID, identifying beneficiaries was provided in each CMS linkable 2008-2010 Medicare DE-SynPUF. DE-SynPUF users can link CMS Linkable 2008-2010 Medicare DE-SynPUFs using this Beneficiary Code, DESYNPUF_ID, as the linking key. However, DESYNPUF_ID was specifically created for DE-SynPUFs and carries no information about the patient or any patient records, and is provided solely for reference and data processing purposes.

Click on the Sample below to be taken to the file download page:

[DESample01](#)

Each of the 20 samples contains eight files – three beneficiary files (one for each year), one inpatient file containing three years of data, one outpatient file containing three years of data, one PDE file containing three years of data, and two carrier files containing three years of data (Carrier 1 and Carrier 2). Because of file size limitations,

a Carrier sample was split into two CSV files. Both CSV files in a sample must be downloaded. Beneficiary data was obtained for each year that the beneficiary enrolled in Medicare. A single Beneficiary sample contains three CSV files, one for each year. Because beneficiary files contain time varying variables like chronic conditions, reimbursement variables, and death, three files (one for each year) were provided to keep the same variable name as in the actual data. All three CSV files in a sample must be downloaded. If the beneficiary dies between 2008 and 2010, the beneficiary will not have any data in years after the beneficiary's death.

Although actual Medicare data are provided by year, De-SynPUFs provided three-year claims files to decrease the number of files users have to download and to take the advantage of multiple year data. Users can easily extract single year claims data.

Click on the file below to begin download:

[DE1.0 Sample 1 2008 Beneficiary Summary File \(ZIP\)](#)
[DE1.0 Sample 1 2008-2010 Carrier Claims 1](#)
[DE1.0 Sample 1 2008-2010 Carrier Claims 2](#)
[DE1.0 Sample 1 2008-2010 Inpatient Claims \(ZIP\)](#)
[DE1.0 Sample 1 2008-2010 Outpatient Claims \(ZIP\)](#)
[DE1.0 Sample 1 2008-2010 Prescription Drug Events](#)
[DE1.0 Sample 1 2009 Beneficiary Summary File \(ZIP\)](#)
[DE1.0 Sample 1 2010 Beneficiary Summary File \(ZIP\)](#)

Section 1 – Problem Statements and Hypothesis

Skills required – R ,Probability, Statistics

Skills acquired – Candidates should be able to articulate hypothesis and create problem statements for machine learning. Candidates are expected to be bayesian thinkers upon completion of this section

Dataset - Only use DE1.0 Sample 1 files

Questions -

L1. Articulate and create workflow of your understanding and approach to problem given. It should clearly mention the roadmap of how you will build and provide insights to stakeholders. (Points 10)

L2. Formulate 1 hypothesis on this sample data which you would like to test/potentially beneficial to know for targeted stakeholders to validate your solution. (Points 20)

L3. Provide Summary Statistics and inferences about data using statistics. (Points 20)

L4. Create R based dashboard (ie. Shiny R dashboard) with data insights , patient timeline , 3-4 different key metrics. Dashboard should have actionable conclusions , not informative metrics. (Points 50)

Section 2 – Data Engineering

Skills required – Basic R , Basic Python , Data Wrangling , Statistics

Skills acquired – Candidates are expected to be expert in data wrangling , generating insights and should be able to create , understand and interpret summary statistics. Candidates should get familiar to different mechanism to ingest the data/on prem/on cloud

Dataset - DE1.0 Sample 1 – 20 files (as stated below)

[Downloadable-Public-Use-Files Page](#)

CMS 2008-2010 Data Entrepreneurs' Synthetic Public Use File (DE-SynPUF)
DE1.0 Sample 1
DE1.0 Sample 2
DE1.0 Sample 3
DE1.0 Sample 4
DE1.0 Sample 5
DE1.0 Sample 6
DE1.0 Sample 7
DE1.0 Sample 8
DE1.0 Sample 9
DE1.0 Sample 10
DE1.0 Sample 11
DE1.0 Sample 12
DE1.0 Sample 13
DE1.0 Sample 14
DE1.0 Sample 15
DE1.0 Sample 16
DE1.0 Sample 17
DE1.0 Sample 18
DE1.0 Sample 19
DE1.0 Sample 20

Problems -

- L1.** Use R/Python efficient method to read all sample files properly , join them (by method stated in user manual or by understanding metadata) and test them by generating summary statistics. (Points 10)
- L2.** Ingest the all 20 samples in any cloud azure/gcp/aws environment , merge them to create generic cohort/dataset for basic modelling. (Points 20)
- L3.** Ingest the all 20 samples in any spark environment , merge them to create generic cohort/dataset for basic modelling. (Points 20)
- L4.** Create complete spark based ETL pipeline using python/pyspark. (use spark standalone/cloud version). Ingest all 20 samples files in spark , merge them using standard primary key/joining criteria , use sparkSQL to run select query on all required data features , save the file as json and store in mysql/bigquery for visualization. (Points 50)

**spark session will be conducted by training team*

Section 3 – Feature Engineering & Insights

Skills required – R/Python

Skills acquired – Candidates are expected to be expert in data wrangling , generating insights and should be able to create , understand and interpret summary statistics. Candidates should get familiar to different mechanism to ingest the data/on prem/on cloud

Dataset - DE1.0 Sample 1 – 20 files (as stated in section 2 above) [Downloadable-Public-Use-Files Page](#)

Questions -

- L1.** Identifying Beneficiaries Enrolled in Different Time Periods (Points 10)
- L2.** Find beneficiaries who enrolled in all three years and had at least one inpatient claim from 2008 to 2010 (Points 20)
- L3.** Number of Claims per Beneficiary by Service Type Over Three Years (Points 20)
- L4.** Create following data tables with required formatting as follows – (Points 50)
- Create Demography Distribution table

Performance-grid

Sex
Male
Female
Race/Ethnicity
White
Black
Other/Hispanic
Year of Birth
Pre-1924
1924–1928
1929–1933
1934–1938
1939–1943
Post-1943

- Reimbursement by Source by Year

	2008 <i>DE-SynPUF</i> Mean	2008 Mean ¹	2009 <i>DE-SynPUF</i> Mean	2009 Mean ¹	2010 <i>DE-SynPUF</i> Mean	2010 Mean ¹
Inpatient						
Total	\$2,544	\$2,900	\$2,519	\$3,000	\$1,441	\$3,100
Medicare paid	\$2,194	\$2,500	\$2,177	\$2,700	\$1,244	\$2,700
Beneficiary paid	\$247	\$200	\$248	\$200	\$145	\$200
3 rd party paid	\$103	\$100	\$94	\$100	\$52	\$100
Outpatient						
Total	\$846	\$1,100	\$1,028	\$1,200	\$580	\$1,300
Medicare paid	\$624	\$800	\$765	\$900	\$434	\$1,000
Beneficiary paid	\$197	\$300	\$234	\$300	\$131	\$300
3 rd party paid	\$25	*	\$29	*	\$15	*
Carrier						
Total	\$1,536	\$2,100	\$1,734	\$2,300	\$1,100	\$2,400
Medicare paid	\$1,172	\$1,600	\$1,338	\$1,800	\$848	\$1,800
Beneficiary paid	\$346	\$500	\$375	\$500	\$239	\$500
3 rd party paid	\$19	*	\$21	*	\$13	*
PDE²						
Total	\$1,965	\$3,200	\$1,725	\$3,300	\$1,192	\$3,400
Medicare paid	\$55	\$100	\$56	\$100	\$57	\$100
Beneficiary paid	\$10	*	\$10	*	\$10	*

Section 4 – Predictive Modelling & Evaluation

Skills required – Advanced R/Python and Predictive Modelling

Skills acquired – Candidates are expected to know basics for model building and different algorithms to train basic model.

Datasets - DE1.0 Sample 1 – 20 files (as stated in section 2 above) [Downloadable-Public-Use-Files Page](#)

Questions -

L1. Clearly define features & Outcome for modelling for hospital readmissions and train basic model. (Points 10)

L2. Train binary classification model , mention contributing features and their importance for predictions. Results of logistics regression model should also be displayed as follows (Points 20)

Example - Results of a logistic regression for modeling

	Odds ratio	95 % Confidence interval (lower/upper)	p value
Age (years)	1.01	1.01/1.02	<0.01
SOFA score	1.09	1.05/1.13	<0.01

L3. Please articulate the following business questions based on findings - (Points 20)

- Where do providers focus on improving the efficacy of their system?
- What's cost saving and projection for next few years if providers utilize these models?

- In management of healthcare institutions like hospitals, should they focus on efficacy? Or should we focus on effectiveness of services?
- How providers can improve their care pathways if they use this model?

L4. Train any binary classification model by following these steps (Points 50) –

1. Apply Any dimensionality reduction technique (PCA ,tsne,UMAP)
2. Apply any unsupervised clustering technique
3. Train binary classification model
4. Tune the hyperparameters as required for algorithms
5. Identify the best threshold and reasoning

Upon completion please answer the following

1. Can providers focus on cluster of patients to improve their readmissions rate rather than focusing on each and every patient? Who could they be?
2. Please design the risk metrics basis of which provider can take decision as to who all cluster of patients may have higher likely to be readmitted?
3. How likely the patient is going to be readmitted if model has tagged them to be high likely ? How do you define and compare them with true readmissions ?
4. Report model performance on following KPIs (AUC, AUCPR, Sensitivity, Specificity, PPV,NPV) and how provider can take decision based on these metrics?
5. Can you run the model for next 5 years based on the data you have used to train? What all implications and how would provider be assured model is good enough to put it in real world ?
6. Calculate P(True Positive)
7. Comment on model drift

Section 5 – Model Operationalization

Skills required – Advanced Python and Predictive Modelling

Skills acquired – Candidates are expected to formulate MLOps pipeline upon completion of this section.

Datasets – All DE1.0 Sample (20) shall be utilized , ingested and used

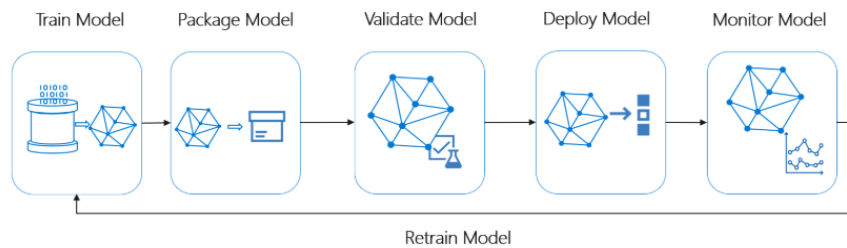
Questions :

L1. Create architecture how to operationalize any model and define operationalization in internal/external environment.

L2. Use any trained model/create new model using R and dockerize the model and expose end point using plumber API. Reference [using-docker-to-deploy-an-r-plumber-api](#)

L3. Use any trained model/create new model , save it as pickle file format and expose end points as rest api. Reference [designing-a-restful-api-with-python-and-flask](#)

L4. Use any cloud environment (AWS/Azure/GCP) to build pipeline for train/package/validate/deploy/monitor model and retraining as well. You can pick up any candidate model which you have built to showcase these 6 sections.



References

<https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs>



Performance-grid