

Business Problem for Midterm Exam – Fall 2023

BIA-810D

October 16th 2023

General Instructions:

You'll use the dataset from "Mid-Term_Project_Material.html/.ipynb" (Syntegra datasets - Medicare CCLF Claims), that have been cleaned and prepared to analyze and address the key business questions below. Extra Credit for building up on these key business questions (i.e. create additional ones) and answering them. Please submit ONE Jupyter Notebook solution on CANVAS.

See sample submission (Udacity Capstone):

https://github.com/divya-nk/udacity-ml-nanodegree-projects/blob/master/boston_housing/boston_housing.ipynb

Business Problem Statement:

The Cardio Vascular Metabolic (CVM) diseases are a cluster of conditions and risk factors that increase the likelihood of developing heart disease, stroke, and/or type 2 diabetes. It is characterized by a combination of several interconnected metabolic abnormalities, including Obesity, Insulin Resistance, High Blood Pressure (Hypertension), Dyslipidemia and Elevated Fasting Blood Sugar. The presence of three or more of these factors is typically used to diagnose this metabolic syndrome. Most common CVM diseases and conditions include heart diseases like coronary artery disease, Stroke, Type 2 Diabetes, Atherosclerosis, Non-Alcoholic Fatty Liver Disease (NAFLD), Chronic Kidney Disease (CKD), Sleep Apnea, etc. Some popular brands in the market for CVM diseases are Xarelto, Toujeo, Lantus, Eliquis, Jardiance, Crestor, Metoprolol, Plavix, etc. The leadership of a Pharmaceutical Company is seeking to gain insights into the trends and landscape of medical procedures conducted in hospitals and clinics. As a member of the Commercial Analytics team, your role is to furnish the leadership with valuable insights, addressing the following Key Business Questions (KBQs) that will inform the decision-making and strategy development of the CVM Sales & Marketing leadership team.

You will need to analyze the Syntegra Medicare CCLF Claims dataset to help answer the following Key Business Questions (Total of 100 points):

1. Based on the trends for the share of CVM claims as a percentage of total claims over the years 2016 through 2018, what are some business insights you can gather? What are some additional analyses you could do based on these trends?
 - a. Hints:
 - i. Calculate the share of CVM claims from years 2016-2018 using top 100 HCPCS/CPT codes
 - ii. Reference the Medicare Analysis discussed during class (Top Procedure + Analysis #1)
 - iii. CVM includes HCPCS/CPT codes related to Cardiac field/procedures

- iv. Think about business insights in terms of key decisions that sales and marketing leadership needs to make, e.g. sales force, sales force tactics (how to word the messaging sent out to HCPs), patient experience
 - b. Expected Technical Output:
 - i. A 100% stacked bar graph
 - ii. Each bar represents a year
 - iii. Each bar consists of percentage of CVM and non-CVM claims
2. Evaluate the HCP behavior in context of claim volume from 2016-2018. How many HCPs are submitting 1 CVM claim; how many HCPs are associated with more than 10 claims, etc.? Once you perform this analysis, explain how this trend can influence the sales force deployment. That is, how would you segment the HCPs and how would you allocate In-Person (sales force) vs Non-Personal Promotions (NPP, i.e. Emails, Social Media, Digital etc.) efforts?
- a. Hints:
 - i. Find the healthcare provider claim count distribution for the CVM claims identified above in increments of calendar quarter
 - ii. Reference the Medicare Analysis discussed during class (Analysis #3)
 - iii. Each healthcare provider has a unique NPI ID; note that Medicare datasets may have multiple columns for NPI IDs including ones for the facility and operating providers that should be not accounted for
 - iv. Use the following segmentation strategy:
 1. Segment #1. Disease Aware (HCPs with 1 CVM claim)
 2. Segment #2. Trialists (HCPs with 2-4 CVM claims)
 3. Segment #3. Rising Stars (HCPs with 5-9 CVM claims)
 4. Segment #4. High-Volume Prescribers (HCPs with 10+ CVM claims)
 5. How would you determine which sales tactic is best for which segment?
 - b. Expected Technical Output:
 - i. A stacked Bar Chart
 - ii. Each bar represents a calendar year (e.g. 2016)
 - iii. Each section in the stacked bar stands for one segment and its value in the bar is the number of providers for that segment
3. Evaluate the Patient Age demographics in the context of claim volume from 2016-2018. Bucket the patients into groups based on their age and explain the trends. How would you position the Marketing Budgets and the Promotions with respect to the changing landscape of the CVM claims and the respective patient segments?
- a. Hints:
 - i. Investigate the trend for patient age demographics for the CVM claims in years 2016-2018
 - ii. Utilize the solution for Q2
 - iii. To calculate the patient age, create a patient age column and subtract the year of patient birth date from the year of the claim date so you have the patient age at the time of the claim.
 - iv. Use the following Segments for the patient age:
 1. Segment #1: 18 – 59 (Patients with age from 18 - 59)
 2. Segment #2: 60 - 69 (Patients with age from 60 - 69)
 3. Segment #3: 70 - 79 (Patients with age from 70 - 79)

4. Segment #4: 80+ (Patients with age 80 and above)
- v. Count the number of claims corresponding to the patients within each bucket by year.
- vi. Also, calculate the year-over-year change percentage in the claim volume for each age group.
- vii. Which marketing and promotion strategies would focus on the patients?
- b. Expected Technical Output:
 - i. A stacked Bar Chart
 1. Each bar representing a year (2016 – 2018)
 2. Each section in the stacked bar stands for an age group and its value in the bar is the number of claims from the patients in that age group
 - ii. A table with year-over-year change percentage in the claim volume for each segment

Project Guidelines:

Criteria	Meets Specifications
Scoping the Project	The write-up includes an outline of the steps taken in analyzing the data set during the project. The purpose of the final answers to key business questions (above 4) is made explicit.
Addressing Other Scenarios	The write up includes a logical approach to this project under the following scenarios:
Defending Assumptions	If you are making any assumptions of the underlying scenarios such as change in patient demographics, size of pharmaceutical company, limitations of the dataset(s) – please defend those with clearly stated assumptions
Accessing the Datasets	Original Datasets: <ul style="list-style-type: none"> - Go to https://www.syntegra.io/download-syntegra-data - Select “Medicare CCLF (Claims)” for the dataset to download - Fill out the required fields and download Prepared Dataset: Mid-Term_Project_Material.html/.ipynb
Submission Format	<ul style="list-style-type: none"> - Place all the relevant files in one Github repository (See example: https://github.com/divya-nk/udacity-ml-nanodegree-projects/blob/master/boston_housing/boston_housing.ipynb) - The code should be written in Python (Pandas/Numpy) and the Jupyter notebook should be in .ipynb format - The code written in the Jupyter notebook should be supplemented by background/context details as well as the details surrounding the technical approach taken directly written in the notebook - The solution should have a written set of conclusive observations from the analysis that provide insight to directly answer each business question - The original datasets used for the analysis should be saved in the same folder as the code notebook

Project Execution:

Criteria	Meets Specifications
Proper Project Code and Format	<ul style="list-style-type: none">- The solution and the original datasets are in a public Github repository- The solution is in one Jupyter notebook (.ipynb format)<ul style="list-style-type: none">o The code is written in Pythono The code runs without errorso All scripts within the Jupyter notebook is clean, easy to follow, and well-structuredo The variable and function names use the PEP8 style guidelineso There are adequate comments/written details such that the reader can understand the context and the code logic without additional explanationo The written answers that translate the analytical output into “layman’s terms” provide clear insight for the business questions asked
Data Quality Check	<p>Your analysis should include at least two data quality checks. Examples include:</p> <ul style="list-style-type: none">- Unique rows (No two records have the exact same data)- No null/empty cells for important columns (e.g. no empty cells for primary keys or join predicate)- Valid datatype (e.g. If a column is expected to have dates in a YYYYMMDD format, then all rows have such formatted data)- Expected output matches actual output
Data Visualization	Each business question analyzed and answered should have a graph or chart associated with it. Pay special attention to the visualization (x-axis labeling, y-axis labeling and colors).
Extra Credit	Extra credit is possible for expanding on the project scope – that is, to create additional questions related to the given key business questions and answering them by further analyzing the Medicare data, whether using existing or new output