EXPLORATORY DATA ANALYSIS USING SPARK

HEALTHCARE IN THE UNITED STATES OF AMERICA

ALLISON BLACK
IE MBD 2020

Table of Contents

What: Background

It is known throughout the world that the American healthcare system is expensive, inefficient, and non-transparent. The Affordable Care Act (ACA), introduced by President Barack Obama in 2010, helped guide the United States towards a better healthcare system. Unfortunately, many measures of the ACA have been reversed and one big issue that remains is access to healthcare. Many Americans cannot afford healthcare, and even if they *can* afford medical insurance, it is often difficult to find the right doctor for an individual's healthcare needs. With technology and big data analytics, I believe that healthcare can be improved tremendously. New tools, such as electronic medical records (EMRs), electronic prescriptions, and virtual doctor appointments, can help to improve patients' access to healthcare and the overall satisfaction with the care they receive.

Using the Flights Analysis (2008) Jupyter Notebook as an outline, I will perform exploratory data analysis on a dataset that contains performance information on hospitals and doctors participating in a Merit-Based Incentive Payment System (MIPS) in the United States. Physician Compare is part of Medicare, the American federal health insurance program for people who are 65 years or older. According to the Medicare website, Physician Compare "helps you find and compare clinicians and groups enrolled in Medicare so that you can make informed choices about your healthcare."[i] Having this data available to patients is a good start. However, many more insights can be taken from the raw data that could possibly lead to better service for patients.

One of the main measures that I will analyze is e-prescribing; therefore I will explain further what it is and why it is important:
E-prescribing, or electronic prescribing, is a technology framework that allows physicians and other medical practitioners to write and send prescriptions to a participating pharmacy electronically instead of using handwritten or faxed notes or calling in prescriptions. E-prescribing systems can create and refill prescriptions for

individual patients, manage medications and view patient history, connect to a pharmacy or other drug dispensing site, and integrate with an electronic medical record (EMR) system. The use of a qualified e-prescribing system is required by the US government's Electronic Prescribing Incentive Program, which gives a medical practice up to a 2% reimbursement of its Medicare Part B charges. "A qualified e-prescribing system must be able to transmit prescriptions electronically, warn prescribers about potential allergic reactions and inform physicians about generic alternatives, among other functionality. E-prescribing also reduces the number of prescription errors attributed to bad handwriting or illegible faxes."[ii]

Why: Goal of Analysis

My hypothesis is that there is a correlation between hospitals using technology (e-prescribing, secure messaging, etc.) and patient satisfaction. With the Physician Compare 2017 dataset, I will analyze data that can help me answer three business questions:

1. What is the ratio of hospital measure performance rate by grade?
2. What are the statistics by type of measure (pain assessment and follow up, e-prescribing, secure messaging)?
3. What are the 100 best and worst hospitals by measure performance rate, and which states are the worst hospitals located in?

A secondary hypothesis, possibly for future analysis, is that if doctors are paid according to their performance, the quality of patient care will improve. At the same time, unnecessary, expensive testing will go down, allowing for cost-cutting and lower insurance premiums.

How: Analysis deep dive

What is Spark?

Spark is one of the industry-leading frameworks for modern big data analytics. Therefore, before getting into my analysis, I will answer the question, “What is Spark?”

Spark is a *unified, in-memory, parallel* computing framework for big data.

- Unified: We can use Spark for multiple use cases (batch processing, machine learning, advanced analytics, etc. Spark provides the user with built-in APIs (application programming interface).
- In-memory: Spark leverages the server’s memory to speed up computation.
- Parallel: Spark works on big data clusters and can deal with any amount of data.

Because Spark is such a powerful processing engine, it is the best choice for conducting my analysis. As a physical wellbeing specialist, I am consistently looking into healthcare data. To find a dataset, I went to Kaggle.com and filtered for medium-sized datasets with the keyword, “healthcare.” When I stumbled upon the “Physician Compare” datasets, I was intrigued. These yearly datasets are an inside look into the performance of hospitals and Medicare providers across the United States.

I used the Reference Analysis (Flights 2008 dataset) as a template for my analysis. After many hours into my analysis, I realized an issue with my dataset: the most interesting information was all in one column. A better dataset to answer my business questions would have each Measure Title of interest in a different column.

Cleaning Data

According to Forbes, “Data scientists spend 60% of their time on cleaning and organizing data. Collecting data sets comes second at 19% of their time, meaning data

scientists spend around 80% of their time on preparing and managing data for analysis."[iii] After some initial data cleaning – mostly renaming columns and removing some columns that were not needed for my analysis – my dataset was ready for some analysis. This was not the only time I cleaned my data; after searching for nulls I again cleaned the dataset.

Setup: PySpark Environment, Spark Data Abstraction

Using code from the Reference Analysis Jupyter notebook, I imported findspark, SparkContext, and SparkSession. The user code (in this case, Python) interacts with the Spark session inside the driver process, which is part of the Spark architecture. PySpark is the bridge between Python and Spark; this bridge is made up of multiple functions that I need to import in order to build the bridge. To import my data and set up my DataFrame, I used spark.read to read and organize the Physician Compare 2017 CSV file.

Dataset Metadata Analysis

To get an idea of the size and shape of my dataset, I imported display and Markdown from IPython.display. I was able to see each column and datatype and the number of rows in the DataFrame (73,321 rows). By selecting two random sample rows, I was able to see what kind of values are associated with the columns. From these two samples, I identified entities, metrics, and dimensions. I then divided my columns into two categories: performance-related columns and location-related columns.

```
1 from IPython.display import display, Markdown
2
3 healthDF.printSchema()
4 display(Markdown("This DataFrame has **%d rows**." % healthDF.count())
```

```
root
 |-- Hospital: string (nullable = true)
 |-- Group PAC ID: long (nullable = true)
 |-- State: string (nullable = true)
 |-- ACO PC ID 1: string (nullable = true)
 |-- Measure Code: string (nullable = true)
 |-- Measure Title: string (nullable = true)
 |-- Measure Performance Rate: integer (nullable = true)
 |-- Denominator Count: integer (nullable = true)
 |-- Star Value: integer (nullable = true)
 |-- Five Star Benchmark: integer (nullable = true)
 |-- Reported on PC Live Site: string (nullable = true)
```

Column Profiling

By doing basic profiling on the two column categories, I was better able to understand the dataset. For example, I could check for nulls, see the most frequent and least frequent hospital, etc. A lot of my data analysis depends on the Measure Performance Rate column; this is the score that hospitals receive from patient surveys on different care-related topics. After seeing that there were 20,493 nulls in the Measure Performance Rate column, I cleaned the data by dropping these nulls from my DataFrame.

Before dropping nulls:

```
Checking for nulls on columns Measure Performance Rate, Star Value, Five
Star Benchmark:
+------------------------+----------+-------------------+
|Measure Performance Rate|Star Value|Five Star Benchmark|
+------------------------+----------+-------------------+
|                   20493|     70125|              70125|
+------------------------+----------+-------------------+
```

Code for dropping and displaying nulls in the “Measure Performance Rate” column:

```
# Drop nulls
healthDF = healthDF.dropna(subset=["Measure Performance Rate"])

print("Checking for nulls on columns Measure Performance Rate:")
healthDF.select([count(when(col(c).isNull(),\
        c)).alias(c) for c in ["Measure Performance Rate"]]).show()
```

vAfter dropping nulls:

```
Checking for nulls on columns Measure Performance Rate:
+------------------------+
|Measure Performance Rate|
+------------------------+
|                       0|
+------------------------+
```

Answering the Business Questions

The Measure Performance Ratings are my most important metric in this dataset. Since these ratings range from 0 – 100, I created categories of the ratings and labeled them "grades." The grades are as follows:

- Excellent: 90-100
- Good: 80-89
- Average: 70-79
- Poor: 0-69

By creating a new DataFrame, called gradeDF, I was able to discover the percentage of measure performance ratings by grade.

Grade	Instances	RoundedRatio
1. Excellent	21803	41.27
2. Good	4679	8.86
3. Average	3107	5.88
4. Poor	4311	8.16
5. Unacceptable	18928	35.83

Of the metrics listed in the Measure Performance Rate column, my topics of interest are: e-prescribing, secure messaging, and pain assessment and follow-up. For my second business question, I analyzed basic statistics on these metrics.

Finally, I analyzed the best and worst hospitals and the states whose hospitals appeared in the "100 worst hospitals" list.

Top 100 Hospitals by average Measure Performance Rate:

Hospital	State	avg(Measure Performance Rate)
SUMMIT PATHOLOGY ...	OH	100.0
AMSOL ANESTHETIST...	IL	100.0
PATHOLOGY LAB ASS...	AL	100.0
MID MICHIGAN INTE...	MI	100.0
PATHOLOGY ASSOCIA...	FL	100.0
RADIOLOGY REGIONA...	FL	100.0
WEST RIVER ANESTH...	SD	100.0
JELLICO COMMUNITY...	TN	100.0
SOLANO ANESTHESIA...	CA	100.0

Bottom 100 Hospitals by average Measure Performance Rate:

Hospital	State	avg(Measure Performance Rate)
MAGNOLIA EXPRESS ...	MS	0.0
PATRIOT URGENT CA...	MA	0.0
GARY L CURSON PA	FL	0.0
PAJARO VALLEY NEU...	CA	0.0
ALLERGY and ASTHM...	OH	0.0
VALLEY ENT, P.C.	PA	0.0
LELWICA CHIROPRACTIC	MN	0.0
SYRACUSE ENT SURG...	NY	0.0
ORTHOPEDIC INSTIT...	CA	0.0
WESTERN INFECTIOU...	CO	0.0

Number of times states show up in the bottom 100:

State	count
FL	11
CA	8
NY	6
GA	6
IL	6
TX	5
CO	5
WA	4
PA	4
AZ	4

Insights: Conclusions

First of all, this analysis displays how powerful Spark is for analyzing big datasets. When dealing with Big Data, an issue that arises is *what* to do with all of the data being collected and *how* to analyze it. Opening a CSV file with 73,000 rows seems to be a daunting, jumbled mess of strings and integers. However, with Spark DataFrames and some organization, I am able to pull some very interesting insights from this data.

When analyzing data for my first business question, I discovered that most of the Measure Performance Rate scores fell into two categories: Excellent (41.3%) and Unacceptable (35.8%). The three categories in the middle – Good, Average, and Poor – accounted for 8.9%, 5.9%, and 8.2%, respectively. This tells me that over 75% of patients who responded to the survey were either extremely pleased or extremely displeased with the level of care they received. Since the American government is paying for the treatments and services of Medicare providers, the government should be aware that over one third of these services come with an “Unacceptable” grade.

My next business question aimed to uncover information about three important metrics: E-Prescribing, Secure Messaging, and Pain Assessment & Follow-Up. Of the three, E-Prescribing has the highest Measure Performance Rating mean at 84.5; this is a positive sign for the future of e-medicine. However, the lowest Measure Performance Rating of the three is Secure Messaging, with a mean of only 18.1. This is disappointing because if going to the doctor’s office or hospital is a problem for the patient, Secure

Messaging can be a good alternative. If a patient needs a prescription refill or has a question about symptoms, secure messaging with a doctor can save everyone time and money. The government needs to be aware that this is an area that needs vast improvement. Finally, Pain Assessment and Follow-Up has a mean of 69.0, which is not good as a mean, but the fact that the third quartile is at 100 is a good sign. For a patient to fully recover and proceed towards good health, Pain Assessment and Follow-Up is key. When a patient leaves the hospital, they might be in pain and could be tired and confused. Any instructions given to the patient might not be remembered correctly, so it is imperative that hospitals improve their patient follow-up calls and messages. By following the correct post-hospital instructions, patients will heal faster and improve their pain-assessment score as well.

Finally, after I found the top and bottom 100 hospitals by Measure Performance Rate, I was able to see how divided the scores were. The distribution, which can also be seen in my grade DataFrame, is clearly weighed towards the top and bottom of the scale. My final objective was to list the states that appear multiple times in the “Bottom 100 Hospitals by average Measure Performance Rate.” The fact that the federal government, and therefore American taxpayers, are paying for these services, means that more attention needs to be brought to those providers who are not delivering quality results. Florida, where around 20% of the population is 65 years or older and therefore qualifies for Medicare, shows up 11 times in the bottom 100 hospitals. Audits, or other checks, may need to be done in these cases.

Link to dataset:

https://data.medicare.gov/data/physician-compare

Link to PySpark documentation:

https://spark.apache.org/docs/latest/api/python/index.html

Link to video presentation:

https://www.canva.com/design/DAEIuOVW9q8/F-wXCcK4hXhoMbH3KNE9Kw/view?utm_content=DAEIuOVW9q8&utm_campaign=designshare&utm_medium=link&utm_source=recording_view

References:

[i] https://www.medicare.gov/physiciancompare/#about/aboutphysiciancompare

[ii] https://searchhealthit.techtarget.com/definition/e-prescribing

[iii] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#:~:text=Data%20scientists%20spend%2060%25%20of,and%20managing%20data%20for%20analysis.