Data Science (COMP 261)

Topics Title: **Analysis of Medicare Health Insurance Data and estimate the health care benefits for low income people**

Prepared By: Syeduzaman Khan

Abstract

US Healthcare is becoming a bigger concern due to its overpriced and inefficient mess. The cost of healthcare cost is rising day by day that makes the situation worst for health care beneficiaries as well as government. Federal government pays its 15% of national budget in Medicare sector. The situation becomes worst for the Medicare beneficiaries when federal government announces the policy for Medicare budget. Therefore, the analysis of Medicare data using Data Science knowledge will provide sustainable solution for beneficiaries and federal government.

Medicare maintains all the patient's data including all personal information, disease types, and prescribed drugs. The data set is huge in volume and challenging to analyze. The development of cloud computing and data science make it possible to query in the huge data set and extract inform from data set. Cloud computing, BigQuey, and Python integrate to analysis Medicare data and find a sustainable solution for the raised problem. The main goals are analysis of Medicare data, finding the relation with immigrants and Medicare benefices, cost estimation in state and national wide, proposal of solution that will reduce the federal government cost on Medical and preventive disease control method using the most prescribed data.

Outline

List of Figures

Introduction

Medicare is a US federal government backed national health insurance program that provides services for low income, children, pregnant women, and disabled people. Medicare health insurance benefits covers almost 55 million people across US. Federal government has spent almost $3.5 trillion in 2017 (CMS 2017)

Almost 55 million health insurance beneficiaries make its one of the largest health insurance service providers in the US. Medicare healthcare data is collectively a group of medical patient's data including drugs information. This public dataset was created by the Centers for Medicare & Medicaid Services. The Medicare insurance dataset has huge volume as well as high complexity. Furthermore, its higher variation and velocity makes it a big data problem. The data set contains all information related to patients and the cost for medicine, services and tests.

The key challenges of analyzing Medicare health insurance data include capturing, storing, sharing, and availability of updated public dataset. The way extracting and organizing the data from data set are also a challenging task. The healthcare data is huge by nature and need much space further if we need related data it needs to undergo with analytics to get relevant data. There are several challenges to deal with this data sets. It is large set of data including 14 columns. The complete datasets contain raw data. It means data mining and cleaning will be big challenge for me. The data set may have some missing fields. To acquire insightful information from it requires a hard work.

The rise of big data has brought a new aspect of solving health care problems. By analyzing patients, insurance company can adjust their pricing. In the recent days, US president Donald Trump has made several statements on Medicare benefits. He has claimed that most of the immigrants take benefits from Medicare that comes from tax payer money. Federal government

has spent almost 15% of its annual budget on Medicare benefits in 2017. That is a huge amount of money so that federal government has made a proposal to cut down the Medicare system. If it happens, the Medical beneficiaries including children and disable people will suffer most. Therefore, an extensive study is needed on Medicare health insurance data. A fruitful study may help government to rethink on this issue and take low cost preventive measure to save the spending's on Medicare.

The objectives of this project are to find the relation between immigrant and Medicare beneficiaries, to identify the govt. cost on per person for Medicare medication, and to come up with a predicative proposal for reducing the government cost on Medicare health insurance.

Related Work

Big Data or Data science offers tremendous insights in health insurance domain. Data science plays an important role in achieving predictive analysis in the health insurance data. It also helps to solve issues and challenges arising in the healthcare system and better analysis of huge data paves the way for making rapid advancement in the health care data understanding which includes new diseases, reducing cost, enhancing medicine and prevents the disease outbreak (Sabharwal, Gupta, & Thirunavukkarasu, 2016) .

People in US spend more than 10% of their gross income on healthcare services. They depend 70% to 80% on the services and information provided by Medicare through internet. The data science strategies and capabilities are used for health data analytics, providing eHealth services, to develop systems for the early diagnosis of the diseases, to prevent and control chronic diseases and also to provide privacy and security for personalized treatments (D K Thara, B G Premasudha, V Ravi Ram, & R Suma, 2017).

Machine learning algorithms works on centralized databases. When it comes to process a large volume of data parallel computing may be helpful. Healthcare domain has seen lots of research in this area where researchers tried to train the system and predict the expected outcome for the patient. An effective decision-making system in healthcare domain using the existing machine learning algorithms has been introduced in this paper (Ramesh, Suraj, & Saini, 2016) .

Strategic implications of data analytics of health care sector could help the health care organizations to understand its capabilities and potential benefits. A predictive model has been proposed by Wang et al. for healthcare organizations that will support them seeking to formulate more effective data-driven analytics strategies (Yichuan, Lee, & Byrda, 2018) .

Big Data analytics can revolutionize the healthcare industry. It can improve operational efficiencies, help predict and plan responses to disease epidemics, improve the quality of monitoring of clinical trials, and optimize healthcare spending at all levels from patients to hospital systems to governments. Nambiar et al. has been proposed a big data-based analytics design to improve overall quality in healthcare insurance cost (Nambiar, Sethi, Bhardwaj, & Vargheese, 2013) .

Solution and Scalability Justification

Scalability in data science both combines the hardware and software aspects as well as the people and process aspects. This includes several factors: data volume (number of rows, columns, and overall bytes), algorithm design and implementation for data preparation, and workflow complexity (Project scalability. 2018).

- Cloud computing: When the question comes regarding scalability, my project meets all the above requirements. I have considered about the scalability of the

project and solutions. Medicare data set has huge volume and continuously its growing. To extract fruitful data within this data set, the computation power and cost comes in the scene. Could computing can be a best option to solve the computation power and cost. Google provides free service for the queries of the first month 1 TB per month. Even if it requires to perform query more than 1 TB per month, the pricing policy is minimal. Google's IaaS and SaaS helps me to get rid of extra cost for data storage (Public datasets.2018) . In future, my project will even sustain if the database becomes larger.

- Open Source tools: I have used open source and free tools to complete my project. Medicare data set can be analyzed using legacy SQL or standard SQL queries. The data is accessible via BigQuery UI or classic web UI or BigQuery REST API. Above all mentioned software package is highly enriched and sustainable for Big Data that saves me from thinking over scalability.
- Algorithm: My proposed solution techniques are free from algorithmic complexity. Therefore, my project will be compatible with larger data set.
- Dynamic and Versatile: Dynamic and versatility are two mains integrated of my project. The proposed can be easily transferable to another domain or in big scale. Also, it uses cloud technology that does not bound the project execution and solution in one geographic area.

Overall, I think that my project is scalable and sustainable in bigger scale.

Implementation Details

Medicare data set has been uploaded to Google server in different geographic locations and available to the general public through the Google cloud public dataset program. The public datasets use BigQuery hosts for accessing and integrating for specific projects. There are several data set available in Google public data sets including Medicare, Iris etc. For my project, I have used Medicare data set.

The data set summarizes the utilization and payments for procedures, services, and prescription drugs provided to Medicare beneficiaries by specific inpatient and outpatient hospitals, physicians, and other suppliers. Also, it includes the following data: common inpatient and outpatient services, all physician and other supplier procedures and services, all prescriptions. Providers determine what they will charge for items, services, and procedures provided to patients and these charges are the amount that providers bill for an item, service, or procedure.

- Software tools: **Cloud computing (IaaS and SaaS), BigQuery, Ipython/ Jupyter Notebook, Python library: Pandas, Numpy, Matplotlib, seaborn, and Plotly**

Medicare data set is maintained by Google and can accessible through BigQueryweb UI, classic web UI command tool or BigQuery REST API using client libraries such as Java, .NET, or Python. BigQuery (written using SQL) a RESTful web service that enables interactive analysis of massively large datasets working in conjunction with Google Storage. BigQuery has been used to run query in data set. The data set has high volume so that I did not download it using RESTful API. The alternate easy solution was to run query using BigQuery. The query execution time was comparatively less than SQL. To create a project in Bigquery public dataset, these are the steps:

create a GCP project, enable billing system (free for 1 TB query per month that was sufficient for my project), and enable BigQuery API. There are two user interface that can be used to access the public datasets: BigQuery web UI and classic web UI (BigQuery.2018) . I have chosen the BigQuery web UI.

Data visualization tools help BigQuery data for visualization and analysis purposes. The best solution is to use BigQuery Python client library and pandas in Ipython notebook/Jupyter notebook to visualize data in the BigQuery sample table. BigQuery Python client library runs the given query and convert the results to a Pandas data frame optionally save the results to a variable and finally display the result.

The results can be exported as csv or another file format. The static graphs were produced using Python matplotlib and seaborn libraries. Plotly was used to plot the interactive graphs.

Results and Discussion

Federal government had claimed they are spending huge money on healthcare to support the low-income US citizens (Michelle Singletary, 2018). But immigrants take the advantages of Medicare and government spending all going in vain (Kenneth T. Walsh, 2018). Fig. 1 and fig. 2 partially supports the government claims. In fig. 1, we have plotted the top five states where immigrants reside most. We have found that California has the largest number of immigrants about 9.8 million. The other top five states are Florida, New York, Texas, and Pennsylvania respectively.

Figure 2 shows the total claims count by state wise. The highest number of insurance benefits had claimed in California nearly 118 million in a year where the highest number of immigrants lived. The other states stand in terms of total claims are Florida, New York, Texas,

and Pennsylvania respectively. That partially supports the government claims on Medicare spending on immigrants.

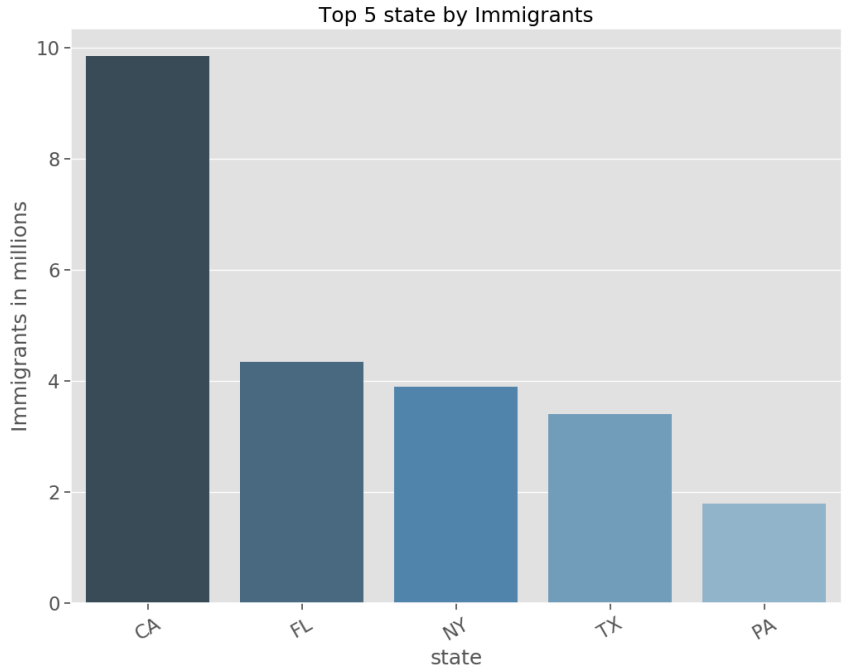

Fig 1: Top 5 state by immigrants' number

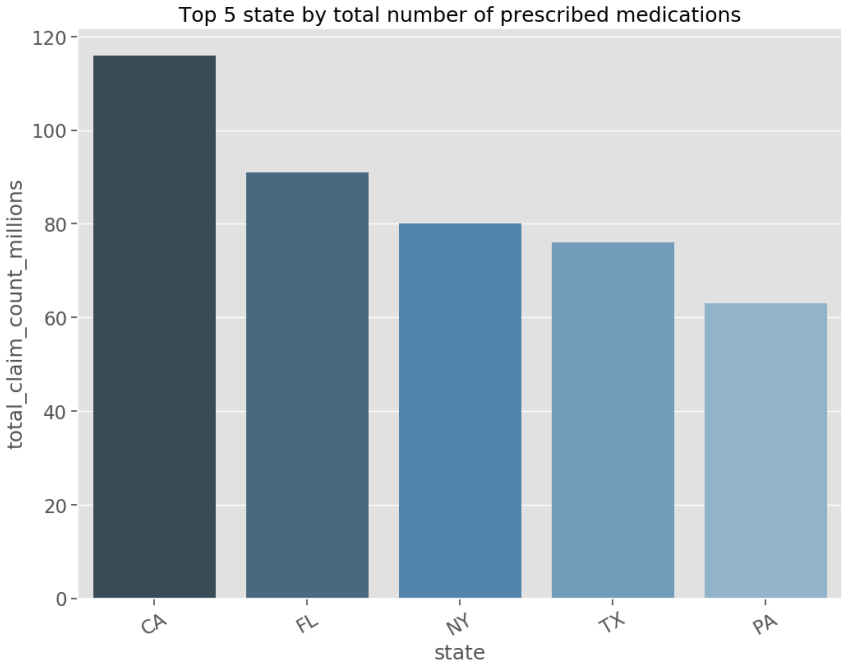

Fig. 2: Top 5 state by total number of prescribed medications

The per person drugs cost shows in figure 3. California stands first in the rank. It spends almost $9.6k on per Medical beneficiaries. New York, Florida, Texas, and Pennsylvania pay almost $7500, $6900, $6400, and $4800 per person. The major portions of the spending go to prescribe the drug namely Levothyroxine Sodium (see fig. 4). It is almost 25% of all prescribed medicine. Levothyroxine Sodium uses to cure hypothyroidism. More precisely it is prescribed for to recover from iodine deficiency. The disease is preventable. Different country has solved this problem by adding iodine with salt.

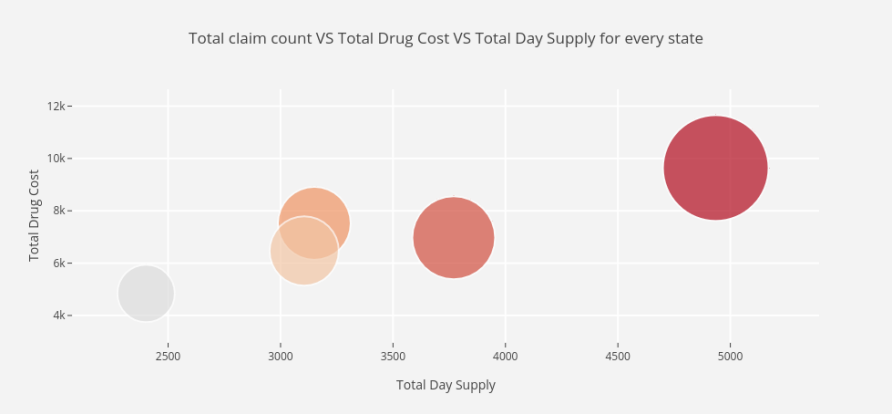

Fig. 3: Total claim vs Total drug cost vs Total day supply for top 5 state (url: click this link)

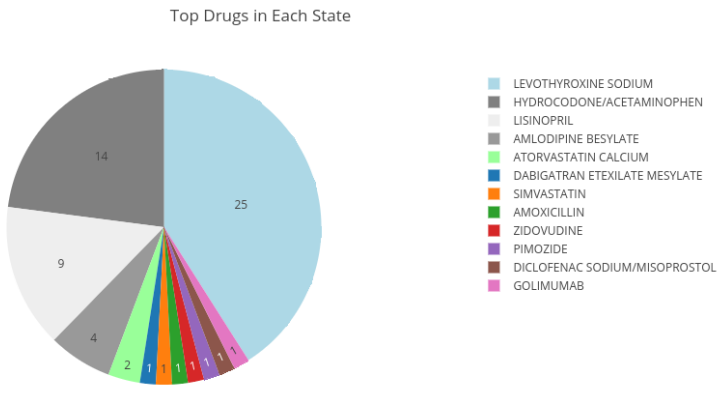

Fig. 4: Top drugs for top 5 state (url: click this link)

Insulin has the highest per unit cost (fig. 5). It is used as prescribed drug for diabetic treatment. Medicare also supports for older people (>65) and costs a huge amount of money.

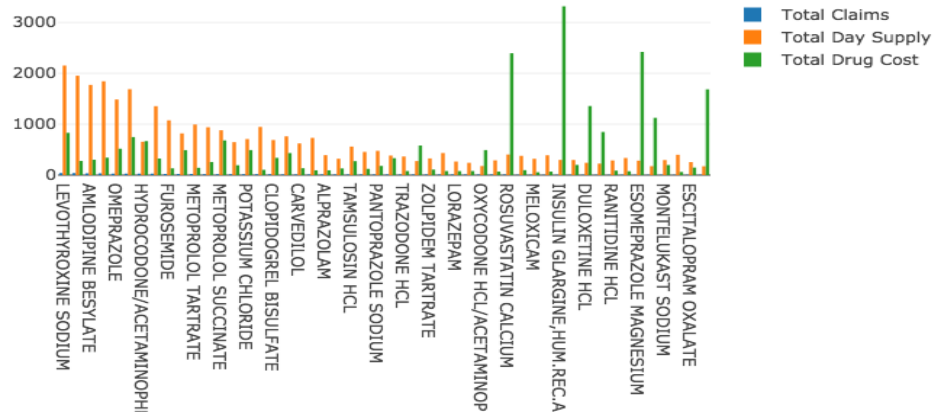

Fig. 5: Total claim vs Total drug cost vs Total day supply in US (url: click this link)

Summary

Data science will really become valuable to healthcare data analysis. The implementation of the project idea provides a critical opportunity to reform the nation's healthcare system. As the federal government moves forward in reducing Medicare budget, the proposed solution can be helpful for making new decision for government policy making. The proposed solution gives an insight of existing disease. So, disease control authority can take initiative to prevent it.

State-wise detailed analyses can be undertaken on the key areas of focus identified in this study so as to gather more in-depth technical and context-specific information. Suitable programs can be designed for each focus area. A similar study can be taken up for the entire health insurance sector including the private health insurance services.

References

BigQuery (2018). Retrieved from https://cloud.google.com/bigquery/docs/visualize-jupyter

CMS (2017). Retrieved from https://www.cms.gov/research-statistics-data-and-systems/statistics-trends-and-reports/nationalhealthexpenddata/nhe-fact-sheet.html

D K Thara, B G Premasudha, V Ravi Ram, & R Suma. (May 2017). Impact of big data in healthcare: A survey in healthcare; Paper presented at the 2nd *International Conference on Contemporary Computing and Informatics (IC3I)*.

Kenneth T. Walsh. (2018). The politics of Medicare and Medicaid, 50 years later. Retrieved from https://www.usnews.com/news/articles/2015/07/30/the-politics-of-medicare-and-medicaid-50-years-later.

Michelle Singletary. (2018). Attention, seniors: Trump's budget is coming for your Medicare benefits; Retrieved from https://www.washingtonpost.com/news/get-there/wp/2018/02/19/attention-seniors-trumps-budget-is-coming-for-your-medicare-benefits/?utm_term=.d6410b283530

Nambiar, R., Sethi, A., Bhardwaj, R., & Vargheese, R. (2013). A look at challenges and opportunities of big data analytics in healthcare. Paper presented at the *IEEE International Conference on Big Data*.

Project scalability (2018). Retrieved from https://blogs.oracle.com/r/data-science-maturity-model-scalability-dimension-part-8

Public datasets. Retrieved from https://cloud.google.com/bigquery/public-data/

Ramesh, D., Suraj, R., & Saini, L. (2016). Big data analytics in healthcare: A survey approach Paper presented at the *International Conference on Contemporary Computing and Informatics (IC3I)*.

Sabharwal, S., Gupta, S., & Thirunavukkarasu, K. ("2016"). Insight of big data analytics in healthcare industry. Paper presented at the *International Conference on Computing, Communication and Automation*.

Yichuan, W., Lee, A., & Byrda, A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Paper presented at the *Technological Forecasting and Social Change*, page 3-13.