MWSUG 2019 - Paper IN-044

The SAS® Enterprise Guide® Process Flow: A Customizable Data Mining Tool in the Search for Healthcare Fraud

Andrea McLain, Cigna Health Insurance, Hartford, CT

ABSTRACT

It is estimated that tens of billions of dollars are lost each year in fraudulent healthcare insurance claims. The implications of this go way beyond financial losses and higher insurance premiums. For instance, many fraud schemes could result in patient exploitation or harm, or the illicit gains could be used in the furtherance of other criminal activities. Health insurance companies utilize data mining and predictive analytics to identify potentially fraudulent claims. Many third party companies create products just for this very purpose, where algorithms are used to flag claims exhibiting some known fraudulent pattern. Products built by these companies are exceptionally helpful in identifying and ultimately stopping and preventing insurance fraud, but many situations call for more, and a means beyond pre-built algorithms are necessary.

This presentation is about one such instance, where a creative, on-the-fly, data mining process was built within a SAS® Enterprise Guide® project to identify potential health insurance fraud in natural disaster scenarios, such as a hurricane or large-scale wildfire. This presentation will detail how an analyst started with millions of insurance claims and then utilized very simplistic analytical methods within a SAS EG project to generate a small list of potentially fraudulent healthcare providers who billed for services they likely could not have rendered due to circumstances surrounding a natural disaster. The SAS skills required to create this process were basic, but speak to the larger concepts of intelligence analysis and data mining in the identification of a criminal pattern.

INTRODUCTION

All health insurance companies likely have a department called the Special Investigations Unit (SIU), which is responsible for mitigating the company's risk to insurance fraud. As an analyst in the SIU, one of your responsibilities is to produce actionable intelligence that the department can use to conduct fraud investigations. Actionable intelligence can come in many forms, but one simple example is a report detailing a fraud scheme and all involved healthcare providers that investigators are not yet aware of. You can transform insurance claims data into actionable intelligence through a process called data mining.

The natural disaster analysis is a data mining model built in SAS EG that has successfully produced actionable intelligence for fraud investigators at Cigna. In general, data mining can be carried out through many different standard process models. The natural disaster analysis is best conveyed within the steps of the Cross Industry Standard Process for Data Mining (CRISP-DM), which is one of the more common process models. CRISP-DM has 6 steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (McCue, 2015). In the first section of this paper, these steps are defined and then explained by examples from the natural disaster analysis. The second section of this paper highlights a case study on the natural disaster analysis for Hurricane Irma. How SAS EG was used to carry out each step is also detailed when relevant.

CRISP-DM & THE NATURAL DISASTER ANALYSIS

STEP 1: BUSINESS UNDERSTANDING

This is the most important phase of the data mining process where you list the overall objectives of the project (McCue, 2015). The main objective of the natural disaster analysis is to identify providers who billed for services they likely could not have rendered due to circumstances surrounding a natural disaster. It was hypothesized that providers already involved in healthcare fraud may not stop for acts of nature, and therefore, meeting this projects objective would, in turn, potentially identify a variety of different fraud schemes. This is the ultimate goal. The first requirement you need to reach this objective is a recent natural disaster. Then, along with claims data, you need information on the geographic areas that the disaster impacted.

STEP 2: DATA UNDERSTANDING

After outlining your objectives, the next step is to collect and explore the data that you will use in the analysis (McCue, 2015). If you do not already have a very thorough understanding of your claims data then you need to develop it here, because not all health insurance claims are relevant to this analysis. To reduce false positives in the results, you want to focus on claims where the rendering provider is a person (as opposed to a clinic, hospital, emergency room, or some large healthcare conglomerate) who physically provides services to patients in an office. After gaining an understanding of your data, take note of the data tables and particular fields that are necessary for the analysis, and make sure you have everything you need.

Also, you only want a selection of claims where the healthcare services were provided in an area impacted by a natural disaster. To get this, you must first define time and geographic parameters of a disaster. You can easily do this with the FEMA Disaster Declarations Summary, which is an open government data source listing all federally declared disasters by county (Fema.gov, 2018). Download the most recent version of the FEMA Disaster Declarations Summary from https://www.fema.gov/openfema-dataset-disaster-declarations-summaries-v1. Open the file, and filter the ‘title’ column to the natural disaster you are using in your analysis. You now have a list of every county impacted by the disaster with an incident start date and an incident end date for each county.

STEP 3: DATA PREPARATION

The goal of the data preparation phase is to create the final data set that will be analyzed in the modeling step (McCue, 2015). This includes pulling and manipulating your claims data so it is set up for analysis. SAS EG makes data preparation a breeze. First, write a Program in SAS to join together all of the necessary claims tables, select the fields you need, and to filter the claims down to those that meet the needs of the project. This will differ significantly from company to company, but the tables you need to join, fields you need to select, and filters you need to apply should have been laid out in the prior step. It may be difficult to match the exact geographic parameters directly in the program, but you can start by filtering down to the state(s) where the disaster occurred. Set the time frame for about a year before the disaster’s earliest begin date through the disaster’s latest end date, which are defined in the FEMA Disaster Declarations Summary.

Next, in the same SAS EG project, drag and drop your FEMA Disaster Declarations Summary into your Process Flow interface to import the disaster data. If you have not already done so, filter this list to only include fields on the natural disaster relevant to the analysis. If your claims data does not include a county field, you can crosswalk the counties from the FEMA data to zip codes. Then, use a SAS EG Query Builder to set up an inner join between the zip codes from the FEMA dataset to the rendering provider zip codes in your program results. The inner join removes all claims where the rendering provider’s zip is not within the geographic parameters defined in the FEMA Disaster Declarations Summary. The results of this join will be referred to as your “Disaster Claims” table.

STEP 4: MODELING

In the modeling phase, select the type of analytical methods that are best suited for the data you are working with and that will lead you towards the goals of the project at hand (McCue, 2015). The analytical methods are necessary to meet the objectives of the natural disaster analysis are quite simplistic, and are easy to do from the Disaster Claims table with a series of SAS EG Query Builders. This step can be broken down into the following 3 categories: zip code claims analysis, provider claims analysis, and peer comparison.

Zip Code Claims Analysis

The goal is to identify small geographic areas, which in this case are zip code zones, where a majority of providers clearly stopped providing healthcare services. County areas can be large, and a natural disaster may impact one side of a county, but not the other. This is why a more granular geographic analysis is performed. Perform the following steps in the order they are listed:

1. Establish Zip Code Norms: Set up a series of SAS EG Query Builders from the Disaster Claims table to calculate the average number of claims per week within each zip code area. Base the average on weeks that occurred about a year before the earliest incident begin date, but not when the disaster occurred. The resulting table will be referred to as your “Zip Code Norms” table.
2. Count Zip Code Disaster Claims: Set up SAS EG Query Builder from the Disaster Claims table to calculate the distinct count of claims per week that occurred within each zip code during the time frame of the natural disaster. Refer to the FEMA Disaster Declarations Summary for the earliest incident begin date and the latest incident end date for the disaster. The resulting table will be referred to as your “Zip Code Weekly Disaster Claims” table.
3. Compare Zip Code Disaster Claims to the Zip Code Norms: Use the Zip code field to join the Zip Code Weekly Disaster Claims table to the Zip Code Norms table. Create a calculated field to determine the percent change between the weekly average claims per zip code and the weekly count of claims during the natural disaster per zip code. Filter the final ‘Zip Code Claims Analysis’ table to only include zip codes where the weekly count of claims during the disaster decreased by more than 75% below the weekly average. This number is flexible.

Provider Claims Analysis

The goal is to separate providers who stopped providing services during the time of the disaster from those who did not. Perform the following steps in the order they are listed:

1. Establish Provider Norms: Set up a series of SAS EG Query Builders from the Disaster Claims table to calculate the average number of claims per week submitted by each healthcare provider. Again, base the average on weeks that occurred about a year before the earliest incident begin date, but not when the disaster occurred. The resulting table will be referred to as your “Provider Norms” table.
2. Count Provider Disaster Claims: Set up a SAS EG Query Builder from the Disaster Claims table to calculate the distinct count of claims submitted by each healthcare provider during the time frame of the natural disaster. The resulting table will be referred to as your “Provider Weekly Disaster Claims” table.
3. Compare Provider Disaster Claims to the Provider Norms: Use a provider key field to join the Provider Weekly Disaster Claims table to the Provider Norms table. Create a calculated field to determine the percent change between the weekly average claims per provider and the weekly count of claims during the natural disaster per provider. Filter the final ‘Provider Claims Analysis’ table to only include providers where the weekly count of claims during the disaster either increased or stayed about the same as their weekly average. You can start by only including providers the percent change between the weekly average and the weekly count during the disaster was greater than or equal to -15%. This number is flexible.

Peer Comparison

The goal is to highlight providers who kept billing in areas where most of the immediately surrounding providers stopped. The percent change between the weekly average claims per zip code and the weekly count of claims during the natural disaster per zip code represents the behaviors of other providers in the same area. Perform the following steps in the order they are listed:

1. Join Zip Code Data to Provider Data: Use an inner join on the zip code fields to join the final Zip Code Analysis table to the Provider Claims Analysis table. You now have a “Final Analysis Results’ table, although the next few steps provide additional information can be helpful when you evaluate your results.
2. Compare Provider Weekly Norms Zip Code Weekly Norms: Use a calculated field to determine the percentage that each providers’ weekly average claims represents within their zip code. Do this by dividing the provider weekly average by the zip code weekly average, and displaying results as a percentage. The point is to be able to say, for any given provider, that that they typically make up x% of all claims in their zip code.

3. Compare Provider Weekly Claims during Disaster to Zip Code Weekly Claims during Disaster: Use a calculated field to determine the percentage that each providers' weekly claim count during the natural disaster represents within their zip code. Do this by dividing the provider weekly claim count during the disaster by the zip code weekly claim count during the disaster, and displaying results as a percentage. The point is to be able to say, for any given provider, that they made up x% of all claims in their zip code during the week(s) the disaster occurred.

You now have a list of providers who may have likely billed for services they could not have rendered due to circumstances surrounding a natural disaster. If the disaster spans multiple weeks, it may be easier for you to create a separate "Final Results" table for each week.

STEP 5: EVALUATION

The goal of the evaluation step determine if your model is meets the objectives set up in the business understanding phase (McCue, 2015). If the disaster spanned multiple weeks, start by conducting additional research on the providers in the final results for the first week. Additional research on a provider is broken down in the following list:

- If relevant, check to see if the provider continued to bill more than their peers during subsequent weeks of the disaster.
- Look at the percentage of weekly claims the provider typically represents within their zip code, and compare it to the percentage of claims the provider represented during the week of the disaster.
- Verify the provider's rendering address and the provider's specialty with other sources.
- Look into the specific types of services that the provider rendered during the disaster.
- Check your case data to see of the provider was ever under investigation.
- If relevant to your company, check to see if the provider's claims have a lot of auto-generated red flags throughout the last year.

If you cannot find a legitimate reason as to why the provider rendered healthcare services during the disaster in an area where most of the other providers stopped, then you need to communicate results to the investigations team. Finding false positives are also helpful, as long as you gain an understanding on how they ended up in your results. This allows you to go back to the data preparation stage and refine your filters to remove them. It is important to refine your filters since this SAS EG process flow can be a template for future natural disaster analyses. The geographic and time parameters will need to be modified per disaster.

STEP 6: DEPLOYMENT

The deployment step is simply where you disseminate your important findings (McCue, 2015). For relevant providers, you can summarize your research in a report, and present findings to members of the investigations team. Make sure the fraud investigators understand the analysis and why the providers are being referred to them.

CASE STUDY

HURRICANE IRMA DISASTER ANALYSIS

Hurricane Irma was a Category 4 hurricane that hit Florida between 9/9/2019 and 9/11/2017. Over 6.5 million Florida residents were ordered to evacuate, and damage estimates were as high as $50 billion (Amadeo, 2019). Many areas of Florida flooded, and winds up to 142mph were recorded (Amadeo, 2019). It was hypothesized that providers already involved in healthcare fraud may not have stopped for Hurricane Irma, so a Hurricane Irma Disaster Analysis was conducted with the goal of identifying providers who kept billing for services they likely could not have rendered due to the impact of Hurricane Irma. The below tables provide mock results from the Hurricane Irma Disaster Analysis:

Provider Name	Zip Code	Provider Claim Count in Week 37	Provider Avg Weekly Claims	Provider % Change from Wkly Avg	Zip Code % Change from Wkly Avg
Dr. Needles	34141	15	15	0%	(91%)
Joe Flooded	34116	8.5	10	(15%)	(77%)
Cash Shark	33178	24	22	9%	(77%)

Table 1. Week 37 Hurricane Irma Disaster Analysis

Provider Name	Zip Code	Provider Typically Represents x% of all Claims in Zip Code	Provider Represents x% of all Claims in Zip Code during Disaster
Dr. Needles	34141	9%	87%
Joe Flooded	34116	8%	40%
Cash Shark	33178	25%	70%

Table 2. Week 37 Hurricane Irma Disaster Analysis Cont'd

All of the columns from Table 1 and Table 2 are included in the final results table in the SAS EG Process Flow. Looking at the first Provider, Dr. Needles, you can glean that she continued with her exact same billing pattern in an area where a majority of other providers stopped rendering services. The number of services she provides typically represents 9% of all claims within zip code 34141, but during Week 37 when Hurricane Irma impacted the area, the services she provided represented 87% of all claims in the same area.

Additional research was performed on this provider. It was discovered that she is an acupuncturist who provided therapeutic services when Hurricane Irma impacted the area. She owns her own acupuncture business, and operates out of just 1 address in zip code 34141. She was never under investigation by the SIU in the past, but many of her recent claims were flagged for upcoding. This research was communicated to the investigation's team, and the SIU opened a case on Dr. Needles.

THE DR NEEDLES INVESTIGATION SUMMARY

Fraud investigators discovered that Dr. Needles was providing a variety of medically unnecessary and experimental procedures; many of which were outside of the scope of her license. A time study also revealed that she was billing for more hours within a day than her office was open. She was also billing for services provided on Sundays when her office is supposedly closed. These fraud patterns were observed throughout the prior 2 years. She is still under investigation.

CONCLUSION

SAS EG is an exceptional tool for creating custom data mining models. The Hurricane Irma Disaster Analysis was quick and easy to set up with a SAS EG Program and SAS EG Query Builders, all within the same process flow. Results were researched the same day the analysis was set up, and actionable intelligence was disseminated to investigators the following day. This has been repeated for many other natural disasters. As displayed through the Dr. Needles case study, the natural disaster analysis is a valuable asset to the SIU.

REFERENCES

Amadeo, Kimberly. 2019. "Hurricane Irma Damage was $50 Billion." Accessed August 15, 2019. https://www.thebalance.com/hurricane-irma-facts-timeline-damage-costs-4150395

McCue, Colleen. 2015. Data Mining and Predictive Analytics Intelligence Gathering and Crime Analysis. Waltham, MA: Butterworth-Heinemann.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Andrea McLain
Cigna Health Insurance
Andrea.mclain@cigna.com