Fargo Health Group: Predictive Analytics for Medical Examinations

Kenda L. Ransom

University of Wisconsin at Green Bay

Abstract

As companies grow and try to get an edge on their competition, data-driven decision making will play an important key role in gaining an advantage. This paper will outline how a business problem is solved using data analytics from the beginning to the end in laymen’s terms. The paper will identify the problem, the key stakeholders, the approach, necessary data collection and cleansing, and model selection criteria. The paper will also explain the ethical implications of this analysis while showing how this approach can be used to solve additional business problems.

Fargo Health Group: Predictive Analytics for Medical Examinations

An often-overlooked aspect of being a data scientist is the ability to communicate the results of an analysis to the stakeholders of the project in a way they will understand and be able to use to make informed decisions. This report will explain how the results of the analysis fits the defined business problem and how the results can be expanded upon to help potentially solve additional business problems that are structured like this one.

# Method

In reading the information presented in the Fargo Health Group Case Study, there is a significant amount of money that can be lost due to inefficiencies in the processing of the disability compensation examinations. A small portion of the lost revenue is tied to the Fargo Health Centers (HC) inability to process the request within 30 days. Per the documentation, if the HC has not processed the examination within 30 days, there is a $200 fine paid to the Regional Office of Health Oversight (Khachatryan, 2014, p.2). Most of the lost revenue comes when the HC sends the request back to the Local Office (LO) because it knows it will not be able to make the 30-day deadline due to staffing concerns. Once the request is returned to the LO, it is outsourced to an Outpatient Clinic (OC) that is not in the Fargo network. This outsourced request costs Fargo $1,250 per request. Also, when the request is sent out of network, there are no time restrictions for the turnaround of the exam (Khachatryan, 2014, p.2).

This potential lost revenue is a business problem that can be solved using a data-driven approach. A potential additional problem not covered in this case study is the idea of lost patients, which directly relates to lost revenue. The Director if QAO, Jay Rubin, understands this business problem and sees the value in using a data-driven approach to solve the problem. According to the Case Study, Mr. Rubin shared his vision with the CEO of Fargo Health Group, Anthony Bryant. With the buy in of upper management, Fargo has requested a data-driven approach be implemented to solve the business problem. In solving this business problem with a data-driven approach, there must always be a focus on the ethical implications of collecting, storing, and using the data. The result of the data-driven approach should be to reduce, if not eliminate, the lost revenue due to not being able to process disability compensation examinations within the Fargo network. If the disability compensation examinations are delivered in a timely manner, patients will be satisfied and not look to leave the Fargo network. The data model should show the HC’s how to appropriately staff for the number of expected examinations. If the HC’s are appropriately staffed based on estimated examinations, the HC’s will be operating at the optimal capacity. In this case, everyone would be used efficiently and there would not be wasted man-hours or over-worked physicians. The output of the data model could also be used by the LO’s to determine which HC a request should be sent to, which will allow the HC to complete the request within the 30-day window.

When using a data-driven approach to solve a business problem, it is imperative to identify the stakeholders and their associated roles within the project. For this project, the RACI formula will be used to identify the key stakeholders. RACI, which stands for **R**esponsible, **A**ccountable, **C**onsulted, and **I**nformed, allows each stakeholder to be properly identified and makes everyone on the project team knowledgeable about who is involved and supporting the project. The RACI breakdown for this project is as follows:

* Responsible – Kenda Ransom (me)
* Accountable – Jay Rubin
* Consulted – Local Offices and Health Centers
* Informed – Anthony Bryant

## Data Analytics Approach

The data-driven approach taken will initially focus on a small data set and one HC for proof of concept. The data set will focus on forecasting the number of heart examinations for 2014 at the Abbeville HC. To forecast the number of heart examinations for 2014, a collection of previous year’s heart examination data is need. The data set provided by Fargo has some available information from January 2006 through December 2013. This is an ample sized data set that will allow a time series model to be created to identify potential trends, cycles, and seasonality aspects that will need to be accounted for in the forecasting. The data set will be inspected for data issues and potential missing values. A method for cleaning and imputing the missing values will be determined after an initial inspection of the collected data. Having a properly cleaned and imputed data set is the key to having a model that will accurately predict the expected number of heart examinations in Abbeville in 2014. Once a filled and cleaned data set is created, two forecasting models will be created to predict the 2014 heart examinations in Abbeville. The two forecasting models will be statistically compared using the MAD and MAPE to select the best model.

## Analysis of Received Data

The structure of the data that collected to build the model was a multi-tabbed Excel Workbook. Each tab contained data about the various HC in different formats. The Explanation of Dataset (“Explanation of Dataset”, 2014) contains information that explains the data in the various tabs. There are numerous data issues found upon the initial inspection of the data set.

In the Abbeville tab, the data is formatted in a structure that can be used for the final data set. The structure of the Abbeville file contained columns for number of heart examinations, year, and month. Within this structure, there was data inconsistencies in the number of heart examinations including “placeholders” and non-numeric for examination values. These values were removed from the data set and were later imputed for.

In the Violet, New Orleans, Lafayette, and Baton Rouge May 2007 tabs, the format of the data was different from that of the Abbeville tab. The data for these four HC’s was in the format of requested HC, type of examination, date, and request ID. From the Explanation of Dataset, the data returned from these HC’s need to be heart related exams that were requested from Abbeville. In the examination type column, a selection of heart related conditions will need to be determined and explained, as the conditions listed in each HC are not the same. Upon the initial inspection of the examination type column, there are inconsistences with naming convention, spelling of examinations, capitalization of examinations, and data format. The date column also contains heart examination requested from Abbeville that are not dated May 2007. These examinations will be considered and imputed where necessary to complete the dataset.

From the Explanation of Dataset document, it states that the December 2013 tab contains Abbeville exams that were routed to other HC (“Explanation of Dataset”, 2014). The December 2013 data is to be used in conjunction with the Heart-Related Condition Codes to determine which examinations are heart related. Once the Abbeville, heart related examinations are determined, they will be added to the final data set.

## Data Cleaning and Imputation Approach

The data will be cleaned and imputed on a year by year basis. This method is chosen to allow for abnormalities and anomalies within the year to be handled appropriately and not affect the data for the other years.

A function will be created to clean and identify the heart condition examinations based on the examination types from the Violet, New Orleans, Lafayette, and Baton Rouge data. After reviewing the listed examination types, nineteen examinations were selected as examinations related to heart conditions.

For this analysis, the examinations labeled “Chest Pain” and “Stress Test” were not included in the heart examinations since other conditions not related to the heart can cause Chest Pains and Stress.

To handle the different date formats in these worksheets, various built-in functions will be used to analysis the different date string formats for the identified heart conditions. For this analysis, it is important to get the correct month and year from the converted date format since the final data set will be based upon month and year. Once the heart examinations are determined for the specific date formats, they will be summed from each HC and imputed as May 2007 data. The Violet and New Orleans worksheets also has data associated with May 2013. Those heart examinations will also be summed and imputed as May 2013 data.

Using the Explanation of Dataset and a vector of heart condition codes that will be derived from the Heart-related Condition Code tab, the Abbeville heart examinations will be identified. The sum of the Abbeville Heart Examinations will be the data that is imputed for December 2013.

After cleansing and imputing the Excel data, out of ninety-six needed rows of data, eleven were missing, but many more rows still needed additional imputation. The overall approach to the data imputation is to use the MICE package in R, select the imputation method based on what data is currently available for each year, and perform five imputations. The data scientist’s will make a judgement call based on the five imputation values to determine how to select the appropriate value to fill in for the missing data. Data will need to be imputed for 2006, 2008, part of 2009, part of 2010, and 2011. The data is complete for 2007, 2012, and 2013. The imputation needed for the December 2009 to February 2010 data will be handled differently using a different imputation method. For the years 2006, part of 2009, 2010, and 2011, the imputations for the missing data was imputed using the “norm” method, which is a Bayesian linear regression. A decision to use the median value of the five imputed values was chosen for years 2006, part of 2009, 2010, and 2011.

For 2008, one month of data is missing. To impute for the missing month in 2008, it is noted that the Explanation of Dataset states that there is an increase in heart exams at the Abbeville HC in October due to the New Orleans HC being closed for a hurricane (“Explanation of Dataset”, 2014). This information is taken into consideration when selecting the proper imputation method. A decision is made to use a random sample method within MICE and judgement to select the imputation value. The five calculated, random sample imputation values are shown in Table 1:

temp2008.data$imp$Incoming.Examinations

1 2 3 4 5

8 3110 886 875 724 840

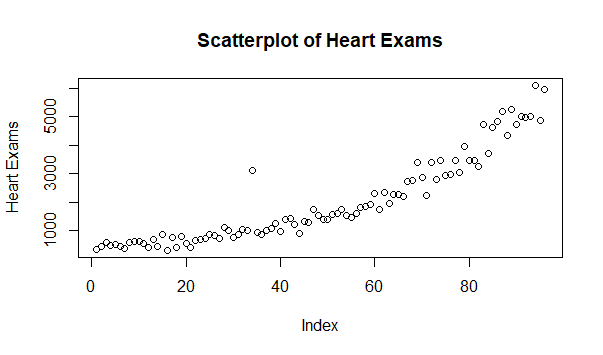
**Table 1: 2008 Heart Examination Imputation Data**

Using a judgement call, it is determined that the first value of 3110 is not a suitable selection as it is the number of examinations that occurred in October 2008, which is the artificially high examination month. Of the remaining four values, value three of 875 will be selected as the imputed value.

The Explanation of Dataset stated that 5,129 heart exams were requested across the Fargo Health between December 2009 and February 2010, but the HC’s that the data came from are not identified. This data will be imputed using some logic based on previous years (2007-2009) data and some general assumptions. The data for 2007, 2008, and 2009 were selected as the data to use for the imputation process because it is two complete years of data were only three months of data needs to be imputed (2007 and 2008) and one year of data were one month of data is imputed (2009).

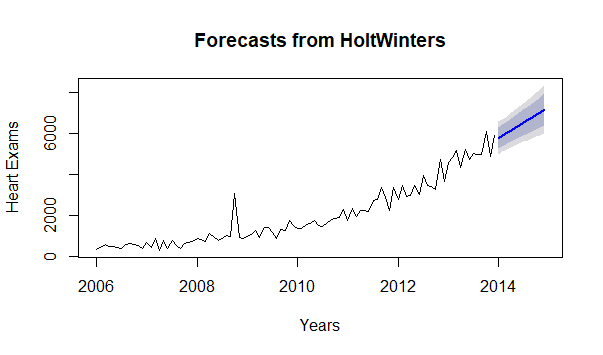
## Forecasting

When creating forecasting models, it is important to look for any trends, cycles, seasons, and irregularities with in the data. Figure 1 is a cleaned and imputed dataset that shows an increasing trend in heart examinations over the years.



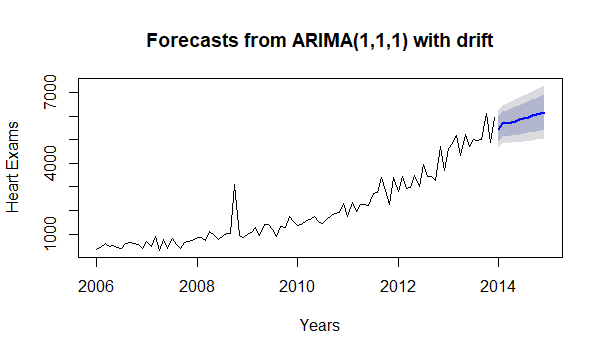
**Figure 1: Cleaned and Imputed Heart Examinations**

Two forecasting models will be created and statistically analyzed to select the best model to predict the 2014 heart examination request at Abbeville. The first model will be the Holt-Winters model which accounts for the increasing trend. Figure 2 is the graphical output of the Holt-Winters model, including the 80% (dark gray) and 95% (light gray) confidence intervals (CI).



**Figure 2: Holt-Winters Forecast with CI**

The second forecasting model is an ARIMA model, which will be auto fit. The auto fit of the model predicted trend and seasonality affects. Figure 3 shows the ARIMA generated forecast with 80% and 95% CI:



**Figure 3: ARIMA Forecast with CI**

Based on the lowest Mean Absolute Deviation, which is the absolute values of errors and the lowest Mean Absolute Percentage Error, which is described as the percent magnitude of error, the ARIMA model gives the best forecast. Table 2 is a comparison of the MAD and MAPE for the two forecasting models.

|  |  |  |
| --- | --- | --- |
|  | MAD | MAPE |
| Holt-Winters | 280 | 20 |
| ARIMA | 256 | 19 |

**Table 2: Forecasting Statistical Comparison**

## Ethical Analysis

The original purpose of collecting the data is for patient medical records at each HC. In the visit to the HC, each patient is given a HIPPA Waiver of Authorization. If the patient signs the waiver, it will allow the medical facility to release the patient’s data to a third-party organization. It was not noted in the provided data was only data of patients who had signed a HIPPA waiver. Using this data for the analysis would be different than the intended purpose of collecting the data if a HIPAA waiver was not signed for each patient in the data set. If the analysis had been done in house, then there would not have been a need to sign a HIPAA waiver.

In creating this forecast, the results can be utilized by many parties for their specific business needs. Fargo Health Group can use the initial results of the model to properly identify what the trend is in heart examinations in the Abbeville HC. Fargo Health Group can also use the results to plan the number of heart physicians that need to be staffed for each month’s predicted examinations. The Fargo Health group employees also benefit from the results of the model as they will be able to see when they can expect the most volume in the office. The Fargo Health Group and the employees can also use the results jointly to plan employee vacations according to the patient load. Public health agencies can use the results of the model to compare with other similar models from different medical facilities to determine what the overall trend in heart examinations is across the population.

Even though numerous parties will have access to the results of the model, it is still important to understand who has ownership of the data, ownership of the model results, and ownership of the insight gained from using the model results. The dataset is owned by Fargo Health Group, but the individual records within the dataset are owned by each patient. It is the responsibility of the project team to treat the data it receives from the Fargo Health Group as though they own the data. The project team must ensure due diligence is taken with the security of the data used for the analysis. At the initial kick-off of the project, there should be a statement which identifies how long the project team gets to retain the raw data. This will help ensure continued security over the dataset. The project team owns the analysis results of the data. The insight that is learned from the results of the analysis and is used to potentially solve additional business problems is owned by the entity that is using the insight.

With all the various entities in touch with the data, the analysis of the data, and the insights gained from the data, it is important to understand the accountability of data/analysis mistakes and what consequences are associated with them. In terms of the raw data, it is ultimately up to the HC’s to ensure the data is properly input into the Fargo Health Group’s data base. The consequence of bad data entry is that it becomes increasingly difficult to gather the necessary, correct information from a query to use in any type of analysis. In terms of the use of data in the analysis, the project team is accountable for ensuring the data is properly cleaned and appropriately imputed and for ensuring the best model to answer the business question is created and utilized for analysis. The consequences of bad data align perfectly to the “garbage in, garbage out” mantra; a bad dataset will only lead to a bad model. If the best model is not chosen to answer the business problem, there is the potential for a bad analysis which can lead to a bad business decision. These risks can end up costing the business more lost revenue that they were trying to recover by initiating the project. If the model and resulting forecast are bad, any entity that wishes to use the analysis and results has the potential to be susceptible to making a bad business decision as well. If an outside entity is using the results for comparison purposes and finds a noticeable difference in the results being compared, they have the option to not use the potential incorrect results in their decision-making scheme.

## Summary and Recommendations

In solving this business problem with a data-driven approach, there are numerous conclusions that can be made. According to the ARIMA forecasting model, the forecast for heart examinations in the Abbeville HC is expected to continue to increase. The forecasted values are located below in Table 3.

Forecasts:

Point Forecast

Jan 2014 5452.034

Feb 2014 5669.661

Mar 2014 5674.158

Apr 2014 5743.677

May 2014 5793.359

Jun 2014 5849.092

Jul 2014 5902.980

Aug 2014 5957.431

Sep 2014 6011.710

Oct 2014 6066.041

Nov 2014 6120.356

Dec 2014 6174.677

**Table 3: 2014 Forecasted Heart Examinations**

Comparing the 2013 Abbeville heart examinations to the 2014 forecasted heart examinations, the predicted increase in number of heart examinations is 15%. This is a significant increase which will require an appropriate increase in the staffing of heart examination physicians to prevent the loss of revenue by not being able to meet the 30-day deadline. With this knowledge, the other HC’s can be aware of the predicted heart examinations at Abbeville and can staff their physicians accordingly if an event were to happen that would require the Abbeville heart examinations to be fulfilled at another HC.

The approximate 15% increase in heart examinations is a significant portion of revenue to the HC and the Fargo Health Group. With the appropriate staffing, there should not be any lost revenue due to inability to complete heart examinations, and the Fargo Health Group should also be able to sustain its patient load, which is positive revenue.

This data-driven approach can be easily spread throughout the Fargo Health Group to predict numerous types of exams and visits. Before expanding this data-driven approach, it will be necessary to standardize some of the inputs to the dataset to ensure that models are being created with the best dataset while upholding ethical standards. From an ethical perspective, each HC will need to present each patient with a HIPAA waiver and explain what this waiver will allow. It is an easy assumption that some people will have reservations about signing the waiver. A potentially straightforward way to overcome this reservation is to have videos playing in the HC where people give their testimony about how predictive analytics help to identify their medical issue, solve their medical issue, or outright prevent their medical issue. For the sake of the models, it is imperative to have a large, diverse dataset. Some of the data that will need to be standardized include:

* Examination Naming Convention
* Relation of Examination Type to Condition
* Date Format
* Data Input Timeframe

Once the necessary data is standardized, the data-driven approach can be used to predict any type of examination or visit for any Fargo HC. System wide, this will help the Fargo Health Group understand how to appropriately staff it’s HC’s to ensure they are operating at the optimal capacity, which will in turn generate the maximum revenue possible and keep the patients in the Fargo Health system.

References

Khachatryan, D. (2014). Fargo Health Group: Managing the Demand for Medical Examinationss Using Predictive Analytics

Explanation of Dataset. (n.d). [Fact Sheet]*.*