Fargo Health Group Predictive Model Report

Fargo Health Group, although thriving with new business and customers, is running into trouble with additional cost to operate in their day to day business. It has been identified that a likely reason for the extra expenses is the steady influx of disability examinations that are submitted to one of Fargo’s 34 Health Centers. These exams are summited to the Health Centers through one of the Local Offices, also ran by Fargo, after it is decided that the Health Centers itself could not self-adjudicate the request. The additional expenses come into play when the selected Health Center cannot perform the evaluation in a timely manner and either contracts out the work to an external Health Center at an average cost of $1,250 above the cost of performing the work internally. Or Fargo is charged $200 from the Regional Office of Health Oversight once the disability exam request is not completed within 30 days. This has caused Fargo to start an initiative to identify ways to mitigate these costs.

My company has been awarded a contract to conduct a pilot study to identify if Fargo can accurately predict the demand of incoming exam requests in the coming months. With the assumption that Fargo Health Group cannot improve on the process for performing disability examinations, my company with work to identify through a data subset provided if a predictive model can be produced. As stated within the contract this pilot study will be focused on one if the 34 Health Centers, located in Abbeville, LA and specifically on exams that are related to heart conditions. My firm has created two different predictive models based on this data, and will be recommending the best within this report. Below you will see the steps performed to develop the predictive models, starting with data cleansing.

In initiating this project my firm received 8 data sets including relevant data for the task at hand. The first was an aggregation of the exam requests (related to heart conditions) for each month of each year for the Abbeville Health Center. This was to serve as the primary source of the data. The rest were delivered to be secondary sources that may help with any erroneous or messy data within the first data set. There were some specific cases that were highlighted when I receive the data that I will go into deeper explanation later on.

To start the process of cleaning this data for use in a predictive model I first started by sorting the Abbeville data by the month and year to try and identify any records that would cause problems with the model. The data points that were found are shown in Table 1. The majority of these problems were identified as likely to be data entry error, they are noted with Data Entry in the column named Reason for Selection. Although the majority were identified as data entry problems there were also some additional cases that were not. The first being the three months between December 2009 and February 2010. These months were not recorded individually but between the three months, there were 5129 cases reported. Also, the last two records within Table 1 show that one year Abbeville closed for the month of December and did not accept any disability examinations. The Clinic was also closed in 2013 for the month of December, but I was informed upon receiving the data that it did receive examination but all were rerouted to close by Clinics.

**Table1**

|  |  |  |  |
| --- | --- | --- | --- |
| Incoming Examinations | Year | Month | Reason for Selection |
| \* | 2006 | 3 | Data Entry |
| \* | 2006 | 6 | Data Entry |
| 99999999 | 2008 | 12 | Data Entry |
| Entered by J. f. Williams | 2009 | 5 | Data Entry |
| \* | 2009 | 12 | Not Recorded |
| \* | 2010 | 1 | Not Recorded |
| \* | 2010 | 2 | Not Recorded |
| xx?\*&?/.. | 2010 | 6 | Data Entry |
| 999999999 | 2011 | 1 | Data Entry |
| Closed for Holidays | 2011 | 12 | Closed |
| \* | 2013 | 12 | Rerouted |
| 3110 | 2008 | 10 | Hurricane |

Additional to the Data problem shown in Table 1, five addition problems were found while cleaning the data for the models, they are listed in Table 2.

**Table 2**

|  |  |  |  |
| --- | --- | --- | --- |
| Incoming Examinations | Year | Month | Reason for Selection |
| 107 | 2007 | 5 | Outlier, suspected Rerouted Exams were not included. |
| 3110 | 2008 | 12 | Outlier, caused by a Hurricane closing a nearby center. |
| 4730 | 2013 | 5 | Noted as incomplete, reroutes were not included. |
| 4706 | 2013 | 6 | Noted as incomplete, reroutes were not included. |
| 5000 | 2013 | 7 | Noted as incomplete, reroutes were not included. |

Once all of the data problems were identified, I worked on fixing the problems by inferring logical data points for these records based on the supplemental data and the characteristics shown within the data set. To start I worked on the data points that real numbers could be identified based on the additional data, specifically the 12/2013 record.

This data point was not recorded as all the exams for that month were re-routed to different clinics. One of the data sets called December 2013 Data contained the information needed to identify all the records that were rerouted from Abbeville in the form of a 17-digit string. These string could then be parsed to identify the origin of the requested exam as well as the type of exam to be performed. For this case, I was only conserved with records that originated for Abbeville and that were heart-related. All Abbeville records start with “L839” and end in either “TGU3” or “ROV8”. Additionally, I was given another sheet that listed the Heart Related Condition Codes which were also parsed from the original string and identified within the table. In Table 3 you can see a subset of the original 17-digit string and that string parsed into its differed pieces. These records were then counted to identify the number of records that were rerouted in December 2013 (5933).

**Table 3**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Original Column | First Four | Next3 | Condition Code | Last Four |
| L839ITSRLX001TGU3 | L839 | ITS | RLX001 | TGU3 |
| L839SMARLX001TGU3 | L839 | SMA | RLX001 | TGU3 |
| L839CRJRLX001TGU3 | L839 | CRJ | RLX001 | TGU3 |
| L839BUARLX001TGU3 | L839 | BUA | RLX001 | TGU3 |

Next, I took a look at the three months from December 2009 to February 2010. As stated before, I knew that between these three months there was a total of 5129 exams. To attempt to make the final model more precise I decided it was necessary to allocate these exams between the three months based on the same three months in other years shown within the data. To do this I located the same months in 06-07, 07-08, and 12-13 as there were the only data points with complete data. From here I calculated a percent of the total for each month within each year which can be seen in Table 4. Looking at the data it was apparent that the year 06-07 was an outlier compared to the other two years so that excluded when I then averaged each month’s average being sure that they still totaled up to 100%. I then took the percentages for each month and allocated the 5129 exams respectively based on their percentage (which can also be seen in Table 4).

**Table4**

|  |  |  |  |
| --- | --- | --- | --- |
| Year | Dec | Jan | Feb |
| 06-07 | 0.366964 | 0.355357 | 0.277679 |
| 07-08 | 0.289266 | 0.362619 | 0.348114 |
| 12-13 | 0.281019 | 0.350704 | 0.368277 |
| *09-10 computed* | *0.285143* | *0.356661* | *0.358196* |
| *09-10 computed values* | *1462* | *1829* | *1838* |

Once Identifying those three data points it was time to take a look at May, June, and July of 2013. These data were identified as incomplete when the data was sent and were noted that all exams that were rerouted during these months were not recorded within the primary table. Additionally, in a step later to be described the month of May 2007 look to possibly be an outlier or missing data similar the months mentioned earlier, so the same process was conducted in this month. To fix this problem we once again look to the supplemental data that was included, this time, the data sheets labeled May-2007 Location were of help. These sheets included data that identified all the exams that originated from one Clinic but were performed in another. To Identify the data for each of the months I filtered the columns by date as well as by the original hospital being Abbeville, the records were then counted and added the original numbers listed within the primary dataset. (NOTE: to keep characteristics of the primary data set the same, and since my medical terminology is limited all records were assumed to be related to heart conditions.)

At this point in the data cleaning process, all of the records that were originally identified as not a data entry error or because of the December 2011 closure have been updated. The remaining data points were updated with the same technique. This process was to plot the corresponding months' data within other years to identify any trends that could help me create a data point that was adhering to the data sets characteristics. The resulting charts can be found in Figure 1-6. It was then Identified that all of the months could then be fitted to an exponential trend line with high R2, the lowest being 0.9729. However, there is one exception to this, May (Figure 6), the R2 returned for that month was 0.7684. Upon looking closer to at the graph it was obvious that the 2006 record was likely to blame, which caused me to include it in the previous step of cleansing the data. Once complete the new data point was added to the chart shown in Figure 7 and the exponential curve was included with a new R2 of 0.9861. the equations for these curves were then used to identify suitable records for the missing months. This process also included the month of October 2008 where the Abbeville Center had an abnormally large amount of exams because of a Hurricane in a nearby city. The original value of 3110 was removed and treated as a null. Once this task was complete the data cleansing phase of this project was also completed. (NOTE: in Figure 1-7 the x-axis represents the year, 2006-2013, of the month in the form of 1-8, and the y-axis represents the number of exams)

**Figure 1 Figure 2**

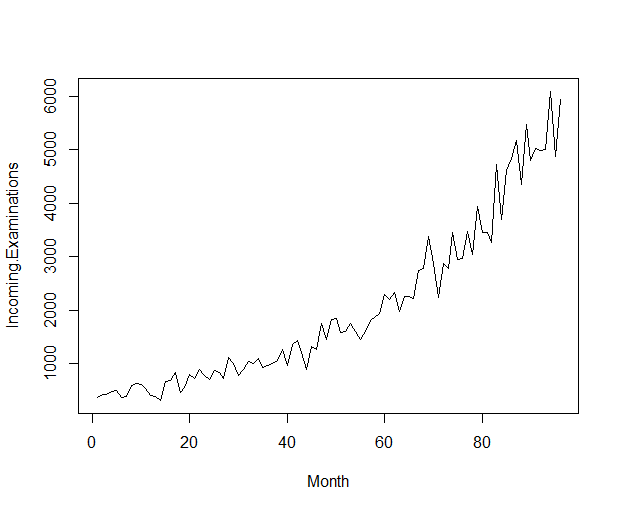
**Figure 3 Figure 4**

**Figure 5** **Figure 6**

**Figure 7**

The next step was to create the two different models, which can be seen in the R file attached with this write-up. My first step for creating the model was to check my data visually to see if there are any inconsistency’s within the data that affect the model being built (see Figure 8), upon doing so it looks to be that my data was in alright fashion with no obvious outliers being shown.

**Figure 8**

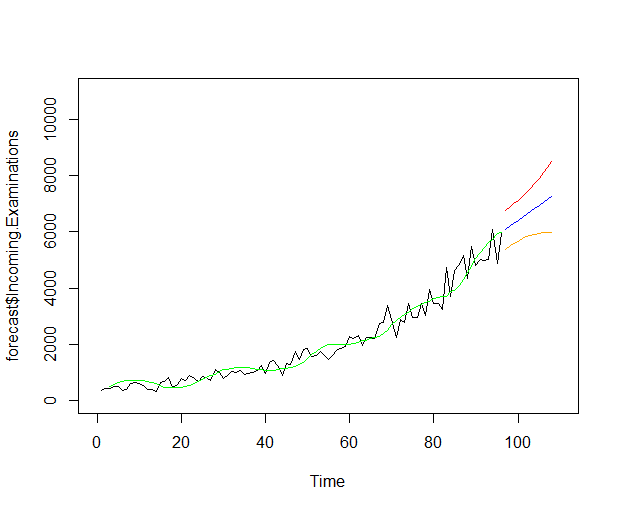


The first model that was created is called a Holt-Winters Exponential Smoothing model, which is an exponential smoothing model that can also take into account seasonal attributes of a time series. Visually speaking, it looking at Figure 8 again it seems no clear seasonal trend. On the other hand, there does seem to be a clear upward trend in the data which we can identify when setting up our forecast model. You can see in the code produced (see Figure 9) that a smoothing constant of 0.06 was used, as the model seemed to return more a more obvious trend for future values. This constant is very easily can be changed and it is recommended if the actual number of exams are not being well represented by the model. Looking back at Figure 9 we also see that beta is set to TRUE and gamma is set to FALSE. These values represent the trend component and the seasonal component respectively. So this model is set to take into account a trend, but to ignore any seasonal components. The final step in setting up this model is telling it how many points to predict and in this model I set it to 12 to represent the next year, this is also true for the second model produced. The returned model including the actual values, the predicted values for the previously recorded values, and the next 12 predicted records with upper and lower bounds can be seen in Figure 10. To evaluate this model, I Computed the MAD, MAPE, and the MSE of the model, the values returned in that order are 272.2,18.5%, and 121784.6.

**Figure 9**



**Figure 10**

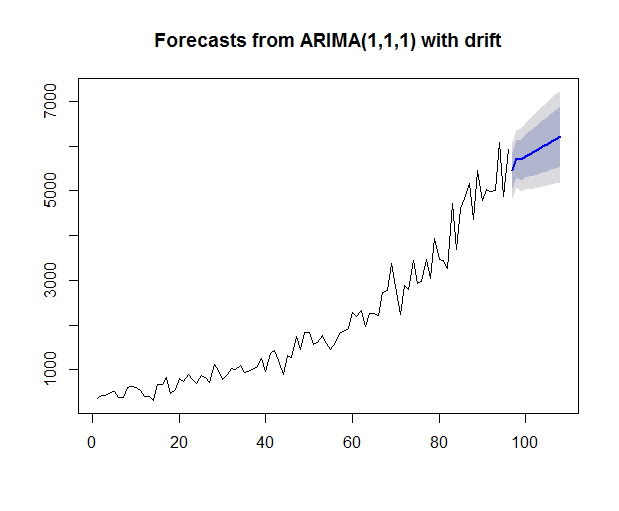


The second Model produced is an ARIMA model, also known as an autoregressive integrated moving average model. This model will mix both a regression to a moving average model, these can be fairly effective for data that has seasonal elements to it. In the R program provided an automated ARIMA model is implemented, you can see an example of the code in Figure 11. This will let R decide the best possible combination of the autoregressive order variable, the differencing order variable, and the moving average order variable on its own. In Figure 12 you can see that the best model returned 1 for all of the variables, you can also see the next 12 months forecasted as well. Now that the model has to be produced we again can evaluate it by computing the values for MAD, MAPE, and MSE. For this model, those values are 220.4, 15.6%, and 48559. Since all of these values are well below the original model produce it is fair to say that this second ARIMA model will be the one recommended for use.

**Figure 11**

****

**Figure 12**



Before this model or any other models is operationalized my firm would like to advise Fargo on the possible ethical implications that could arise. Being in the business of health and medicine it is important that Fargo reviews all HIPAA laws that may be tied to this predictive model. I do believe that the patients represented by these models did give consent for the storage and ownership of these records to Fargo, although maybe not for this specific use. That being said, I would be surprised if this minimal amount of data that was provided to create the models, would violate any HIPAA related laws. If at any point that these models would incorporate any additional or identifiable information related to the patient, and if this data were to be leaked outside of Fargo, the company would be responsible for all legal actions that would arise. All in all, I believe this model will help out all parties involved, including the patients, Fargo Health and its employees as it will now more accurately be able to plan for staffing needs in the future.