Fargo Health Group

Forecasting model report

Shekhar Karmarkar

12 Dec 2019

*Introduction:*

Fargo Health group is hospital group catering to patients for various disciplines. The group provides disability medical examinations. To carry out medical examinations, the group has Local offices, Health centers and quality assessment office (QAO). QAO is responsible for collection of disability examination data from 34 Local offices and subsequent analysis of data. The group has historical data about patients’ examination with information like referring location, date, type of examination, etc.

There is 30 days limit for doing examination, otherwise group has to pay fine to government. Due to lack of adequate physicians, many times group has to either pay fine or refer examination to outpatient clinics, which costs more to the group.

With this background information, I will discuss the problem and the approach to solve it.

**Business problem and necessity of data analytic approach**

The load of examinations and resources like physicians and other staff need to be in proportion. If there is lack of physicians, the patient has to be referred to other locations. This costs money, reduces reputation and it is inconvenience to the patient. On other hand, if many physicians are recruited and no examinations are required, then the resources go waste. There has to be balance between two sides. If by any means we have some idea about future load of examinations, then we can arrange for appropriate resources and save money with improved quality of care.

Recruiting physicians is not an easy task. It requires time and patience. So, getting correct number of future examinations is very critical for overall system. There are two ways to manage this problem. First is by intuition. We can look at current load and judge the future load. But there are always errors, and the decision is subjective.

The better approach is data analytics. In this approach, we take all the data about examinations and analyze it to come to conclusion. We can predict or forecast future demand of examinations more accurate and thus will serve the purpose better. As the model of data analysis is created, we get some idea of future demand. As time goes by, we again collect ongoing data, feeding into the model. The model will be flexible and improve the prediction as more data are collected. This will provide better prediction and serve the purpose.

I got all historical data and analyzed it with creation of prediction models. I created two models and chose better model for prediction. I will go through the process from raw data to predicted values for next 12 months.

When I received the data, it is in raw format. This type of data cannot be used directly to create models. I will go through all the steps to get from raw data to final model.

1. **Cleaning of the data:**

I received data in excel format with multiple worksheets. The main worksheet which will be used for final model is Abbeville, which contains total number of examinations done per month from January 2006 to December 2013. The data contains –

Unrealistic values: like 9999999

Non-numeric values: like text

Missing values,

Incorrect data,

Incomplete data,

I first removed all non-numeric values and unrealistic values. These are replaced by NA or empty cells.

**2. Manage incomplete data**

For incomplete data, I have to gather data from other worksheets, filter them by different criteria and collect relevant figures.

I have collected these four pieces of data to make Abbeville worksheet.

A. For May 2007 data, in addition to 107 examinations, I collected exams from four health centers with these criteria:

-- referred by Abbeville

-- heart related

-- in May 2007 date

The data collected by filtering is --

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Violet | New Orleans | Lafayette | Baton Rouge | Original | Total (May 2007) |
| 281 | 0 | 55 | 52 | 107 | 495 |

B. For Dec 2009 to Feb 2010, there are **5129** cases. I divided this number by three and allocated it to those months.

|  |  |  |
| --- | --- | --- |
| Dec 2009 | Jan 2010 | Feb 2010 |
| 1710 | 1710 | 1709 |

C. For May to July 2013, some cases are in health center worksheets. I gathered those numbers from respective worksheets on these criteria.

-- referred by Abbeville

-- heart related

-- From May to July 2013

The data collected from worksheets is follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Health Center | May 2013 | June 2013 | July 2013 |
| Violet | 0 | 0 | 0 |
| Lafayette | 169 | 19 | 4 |
| Baton Rouge | 145 | 2 | 10 |
| New Orleans | 297 | 0 | 0 |
| Total with original | 5341 | 4727 | 5014 |

All these numbers were added to Abbeville sheet.

D. For December 2013, the data is in separate sheet. I collected data by these criteria.

-- rerouted from Abbeville

-- Heart related condition

The data collected from filtering Dec 2013 sheet gives exams **5933**. It is put in excel sheet of Abbeville.

**3. Data imputation**

After collecting all the relevant data, we have empty cells or NA. By some reasons, the data is not available for some months. The easy way to impute is with mean or median of the dataset. But as this is time series and having trend associated with it, imputation with mean or median is not a good idea. I will use Amelia package to impute the data.

A screenshot of a social media post

Description automatically generatedThis image shows missing values in worksheet. As it can be seen, the red lines are cells with missing data. We have 7 cells empty or NA which is around 7% of total cells. These cells will be filled by imputation using Amelia package.

The image below shows trend of the data after all values are imputed.

A close up of a map

Description automatically generated

**Creating forecasting model**

I will use two methods for forecasting and compare them.

1. Holt- Winters forecast

2. ARIMA model

**Holt- Winters forecast**

This is method of forecasting using exponential smoothing. It accounts for trend and calculates future values. The graph below shows values predicted by this method along with original values. The lines above and below show 95% confidence interval.

A close up of a map

Description automatically generated

As it can be seen from the graph, the model fits quite well to the trend of original values.

**ARIMA forecast**

I have used another model called ARIMA to create forecast values for next 12 months. It isshort for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, so that equation can be used to forecast future values.

The graph below shows values predicted by this method along with original values. The lines above and below show 95% confidence interval.

A screenshot of a cell phone

Description automatically generated

This graph is similar to the one shown for Holt-Winters method.

**Comparing of the models**

Now I will compare two models for predicted values. By comparing two models, we can choose what is better for our use.

There are many parameters to compare the models. Most common used are –

- Mean absolute deviation

-Mean absolute percentage error

-Mean squared error

- Akaike information criterion

I will use MAPE to compare models as this method is easy to interpret and use.

|  |  |  |
| --- | --- | --- |
|  | Holt-Winters | ARIMA |
| MAPE | 43.92 | 16.92 |

From this comparison, ARIMA is better model as the value for MAPE is less for this model.

The mean absolute percentage error (**MAPE**) is a **statistical** measure of how accurate a forecast system is. It measures this accuracy as a percentage and can be calculated as the average absolute percent error for each time period minus actual values divided by actual values.

So, I will choose ARIMA model to create a forecast for next 12 months.

**Further steps**

I want to emphasize that it is necessary to update the model and data. The future values we get is not the end story. As we go ahead and collect more data, we can update the model each month. This will make our forecast much better and will serve the purpose.

**How data analysis helps to solve the problem**

Data analysis helps to address periodic strategic challenges or everyday operational improvements. It helps to coordinate the process of decision making. Experience, market insight and intuition help to make the decisions, but data analysis is science part of the equation. It reduces the stress in the process and make it transparent, and open to challenge.

**Ethics related to Data Analysis**

There are many issues related to ethics when we use data. It becomes more sensitive when the data is very personal like medical examinations. The use of data in medical field is governed by laws like **HIPPA** or Health Insurance Portability and Accountability Act. A major goal of the Privacy Rule is to assure that individuals’ health information is properly protected while allowing the flow of health information needed to provide and promote high quality health care and to protect the public's health and well-being. The Rule strikes a balance that permits important uses of information, while protecting the privacy of people who seek care and healing.

**Original Purpose** of data collection was only record keeping of all examinations and management. We are using data for forecasting which is additional purpose of data usage. This new use is not at all related to the original intended use of data collection.

**Consent** of the patient is not for using the data for this purpose. Patients even do not know that the data will be used for analytics and forecasting purpose. As we have not obtained any consent to use data for this purpose and patient did not have any right to refuse to data collection, it is our responsibility to use this data in prudent manner. The best way to remove patient identification from data is to use de-identification process. In this process, all the demographic data like name, age, sex etc. is removed from data and then used for analysis.

In this way, we use only summary of data and not each individual’s information to do analysis. This is part of **data governance**, where there should be someone who have responsibility to manage the data. If there is data breach, and patient information is leaked, this is serious issue. Not only it is illegal, patients also lose trust in the system.

There are some parameters which can be used to **locate the identity** of the patient by indirect means. Like patient ID number or unique characteristics of patient. We should take care that the data should not reveal identity by any indirect means.

The results of forecasting model should be equitable to all parties like patients, Fargo Health group and all government agencies. The information gained from forecasting models **should not discriminate** against patients. For example, if we forecast that there will be too much load in next month, group can cancel examinations of vulnerable groups. This should not happen.

**Data ownership** is also important aspect. In this case, Fargo health group owns the data. It is responsibility of the group to take care of privacy, security, and access of the data. Fargo health group is accountable for mistakes and unintended consequences in data collection and analysis. As the group has access to all private data related to patients, the group should make sure that the data is used responsibly.

I am thankful to Fargo Health group for opportunity to serve.

**References:**

**1.** <https://www.statisticshowto.datasciencecentral.com/mean-absolute-percentage-error-mape/>

For information of MAPE.

2. HHS.gov for information on HIPAA law and patient privacy