DECEMBER 9, 2018

FARGO HEALTH GROUP

MANAGING THE DEMAND FOR MEDICAL EXAMINATIONS USING PREDICTIVE ANALYTICS

MOHAN RAJENDRAN

BACKGROUND

Fargo Health Group utilizes its Health Centers(HC) to perform disability examination to the patients seeking disability compensations. Once the request is forwarded to the HCs, it is mandated that the HC completes the examination and returns the results within 30 days of receiving the request. But due to lack of physicians HCs often cannot meet the 30-day deadline. Whenever HC foresees such a scenario, they send the request back to Local Office(LO) who will then reroute it to neighboring HCs or to out of network Outpatient Clinics. Despite rerouting, the company ends up losing money owing to the SLA breach due to HCs being understaffed and LO being out of network and don't have any specific SLAs.

Fargo Group came up with a project to implement predictive analytics to forecast HCs incoming requests. This will help them to better manage and implement effective schedule for the examining physicians.

SCOPE

The purpose of this exercise is to develop a predictive model for the Cardio Vascular requests in the Fargo Health Group's Abbeville Health Center.

Data is provided for the Abbeville starting 2006 to end of 2013. The task is to develop two forecasting models and use the best forecast model to forecast next 12 months data.

DATASET DEFINITION

Abbeville, LA contains the incoming monthly aggregate medical examination volume at the Abbeville, LA HC for cardiovascular exams from January 2006 to December 2013. December 2013 data is provided separately, and Heart health condition code is given separately to analyses the file.

In addition to this, data sets are provided for the neighboring HCs which received rerouted request from Abbeville HC.

DATA ISSUES

In addition to outliers and duplicate records, some of the data issues are listed as follows

- Abbeville, LA sheet contains all the cardiovascular request that was receive by the Abbeville HC. Some of the data issues are as follows
- Data in the Abbeville LA sheet has missing values and invalid values.
- Data was incomplete for May 2007, May 2013, June 2013 and July 2013 since the rerouted requests are missing in the totals.
- October 2008 data is a possible outlier since it received more request owing to the closure of a nearby HC
- From December 2009 to February 2010 data is not registered monthly but as an overall total of 5129

- During December 2013 all the requests are routed to neighboring HCs and the data for the same is given in a separate sheet.

DATA CLEANING

For data cleaning, R is used along with excel. Excel is used for creating a dataset around Cardiovascular disease list.

Neighboring HC Data: The first and foremost issue is with the data in the neighboring HCs data sheets.

1. There is no clear definition of what conditions constitutes as the cardio vascular issues in the data
2. Date format is different in each of the neighboring HC sheet
3. There are duplicates in the data
4. Need to create a single data frame

The team with help of little bit of internet research and SME talks was able to identify the list of cardiovascular diseases from the Examination column in the sheets. This list is provided as a separate datasheet. A function CleanData is coded and implemented to resolve the different date format issues and filter the data for cardio vascular diseases and for Original Hospital Location as Abbeville HC.

Once neighboring HC data is cleaned, data is combined to single data frame and duplicates are removed. Having the data in single data frame will help later where we can implement the data cleaning steps once against the combined data frame instead of multiple data sets.

Dec 2013 Data: The data for Dec 2013 is saved with SYSIS which is a combination key that can help us identify the Original Health Center and the Heart Related Medical requests. Requests rerouted from Abbeville have SYSID that starts with "L839" and ends either in "TGU3" or "ROV8". These requests are filtered for heart related disease and the count of the same are updated for Dec 2013.

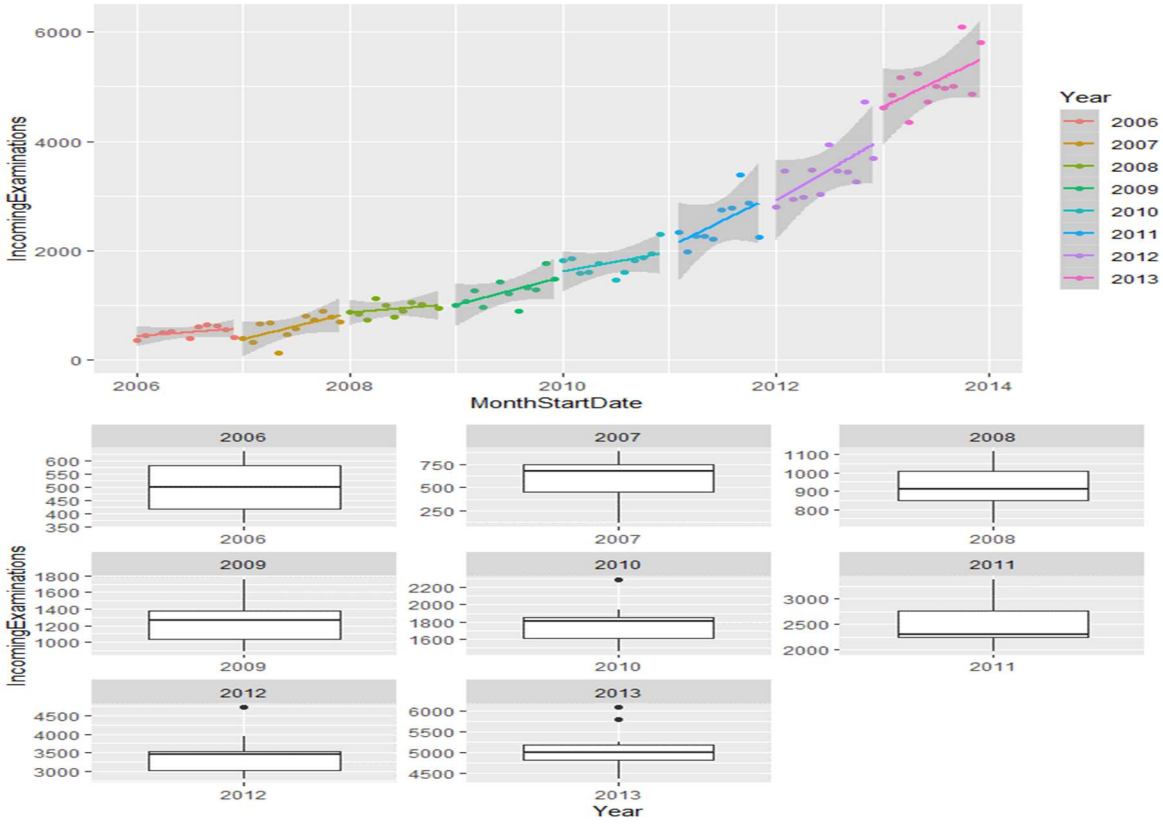
Computing / Correcting Abbeville HC Incoming Examinations Invalid Values: Now that we have both combined Neighboring HC Data and Dec 2013 data we can start computing / correcting some of the invalid entries as follows

1. First, NA is assigned to invalid such as * and other character values and missing values in Incoming Examinations column and then the column is converted to numeric data type.
2. Next, the aggregate value of 5129 is split and assigned for Dec 2009 to Feb 2010 by taking a seasonality ratio by computing average values for these months across 2006-2007, 2007-2008 and 2012-2013 as follows Dec 0.29, Jan 0.35, Feb 0.36. With the help of these averages the data for these months is computed as follows

   Dec 2009 – 1477, Jan 2010 – 1809, Feb 2010 - 1843

3. The next step is to assign NA to the outliers in the data represented by 99999999, 9999999 and for the Oct 2008 when the HC received unusual number of request due to neighbor HC being closed.
4. Lastly, the partial data being displayed for May 2007, May 2013, June 2013, July 2013, Dec 2013 is corrected by adding the existing value with the count of the rerouted entries in the combined Neighboring data frame and the Dec2013 data frame.

Outlier Analysis: The next step in data cleaning is the outlier analysis. In this case charts will be used mainly to identify outliers in the data. The data shows a raising trend over time and outlier analysis needs to keep this under consideration. The dot plot and box plot grouped by year with is as follows:

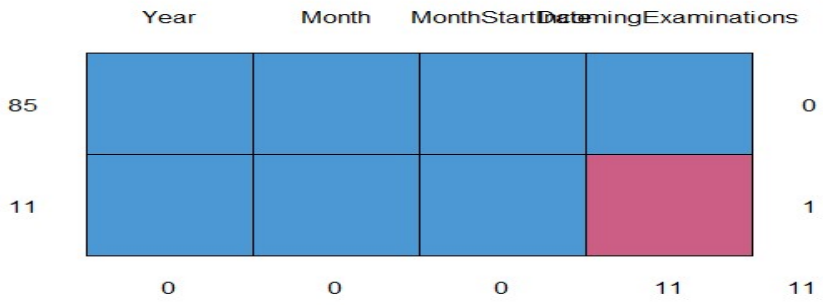


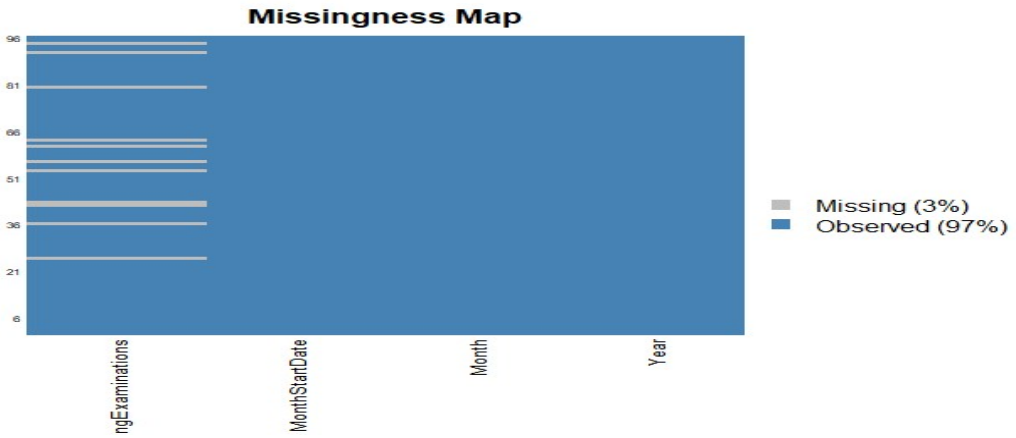
In the dot plot, there are three points May in 2007, August in 2009 and July in 2010 plotted far from the 98% confident interval range. Similarly, the box plot shows outliers in 2012 and 2013.

When considering the rolling date range instead of the yearly date range the outliers identified by the box plot are found to be well within limits. But the three points identified in dot plot are way lower even when considering rolling date ranges. So these three points are marked as outliers and assigned NA.

MISSING DATA IMPUTATION

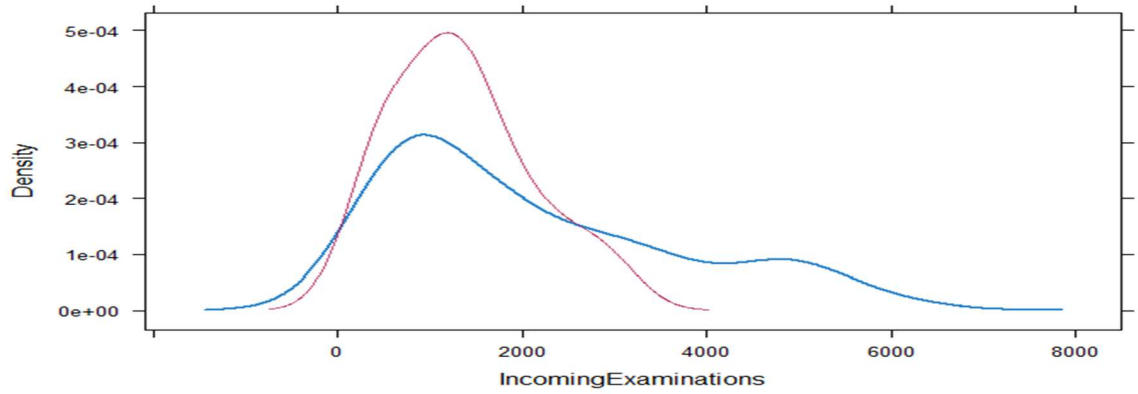
After the data cleanup and outlier analysis the data contains 11 NA values in it. These values form 3% of the data as given by the below plots



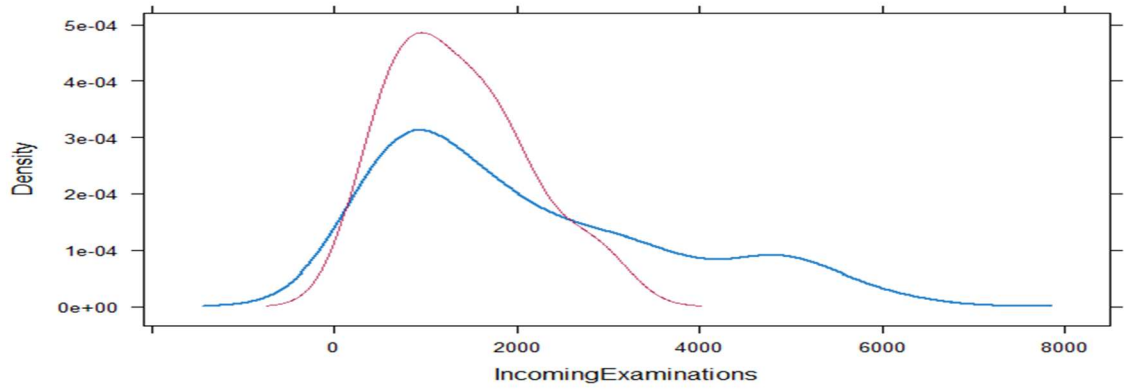


These missing values needs to be imputed so that forecasting model can be developed over the data. Three imputation methods are tried with the data and the density plots are analyzed to find the best fit

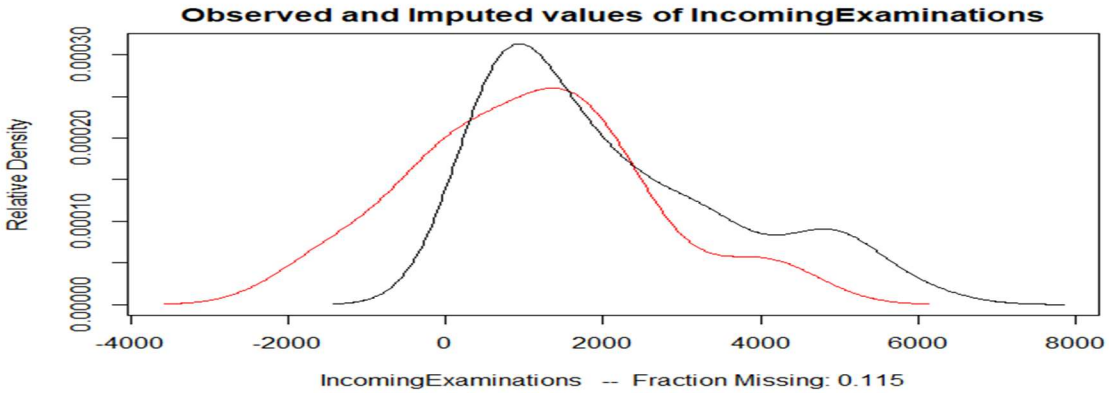
MICE Imputation with Method = pmm - Imputes univariate missing data using predictive mean matching



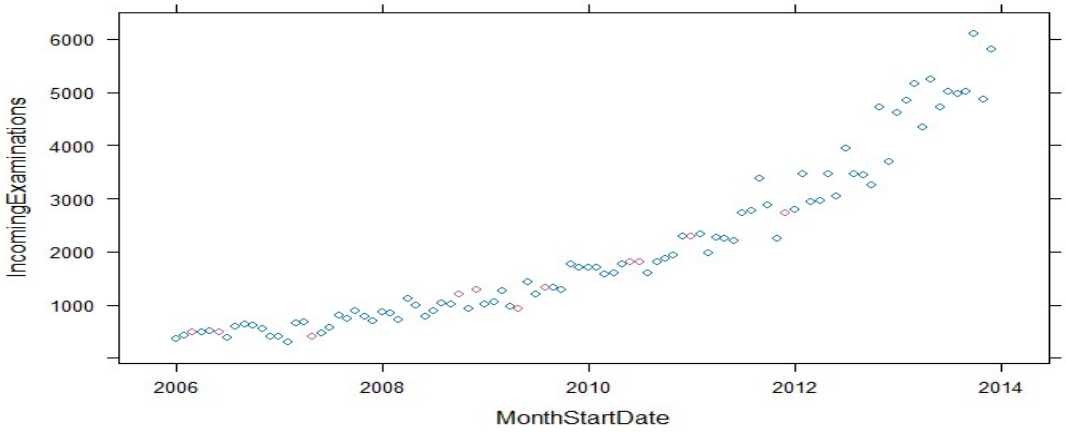
MICE Imputation with Method = cart - Imputes univariate missing data using classification and regression trees.



Amelia Imputation method – Imputes missing data using bootstrapping and Expectation-Maximization algorithm.



Based on the density plots, Mice pmm method is found to impute data that are more aligned with the existing values. Mice pmm method is used to impute the missing values and the xy plot shows the imputed values are in trend with the existing values.

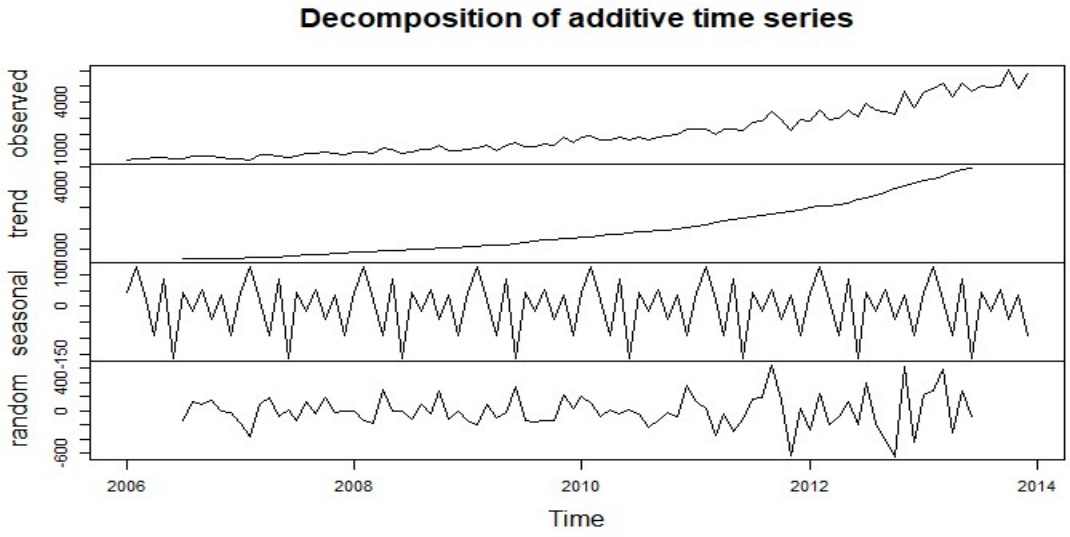


The imputed values are as follows (based on the mice pmm method test run done on 12/09/2018 with seed = 123)

IncomingExaminations	Year	Month
436	2006	3
398	2006	6
613	2007	5
1263	2008	10
962	2008	12
1263	2009	5
1205	2009	8
1604	2010	6
1808	2010	7
2262	2011	1
2869	2011	12

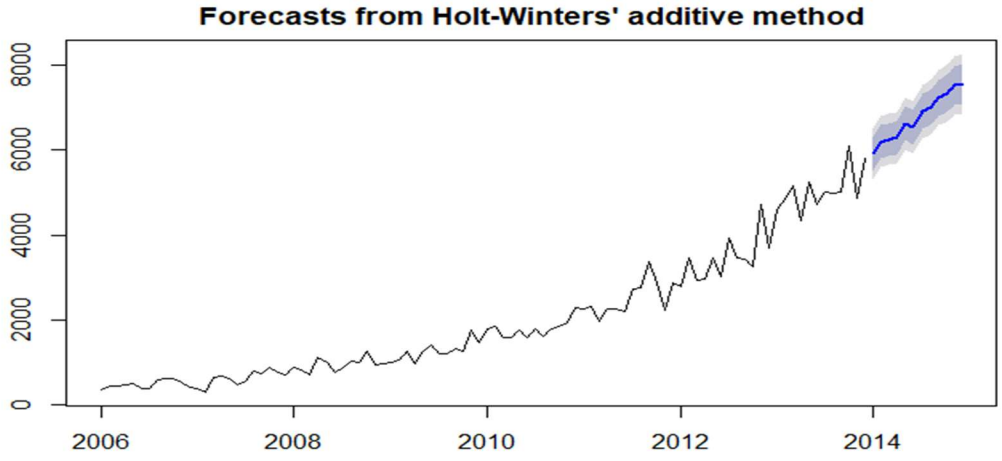
FORECASTING

The data is first analyzed by decomposing it into seasonal, trend and random components. There is a clear seasonal component in the data and hence the forecast model should be able to handle seasonality.



Based on this, it is decided to perform forecasting using Holt-Winters Exponential Smoothing method and ARIMA method. Once the models are developed, we will calculate the accuracy and AIC of the model. Accuracy gives RMSE, MAE, MPE, MAPE which is compared between the two models and AIC

Holt-Winters Exponential Smoothing is used to describe time series with increasing or decreasing trend and seasonality. The model is implemented and the same is plotted as follows

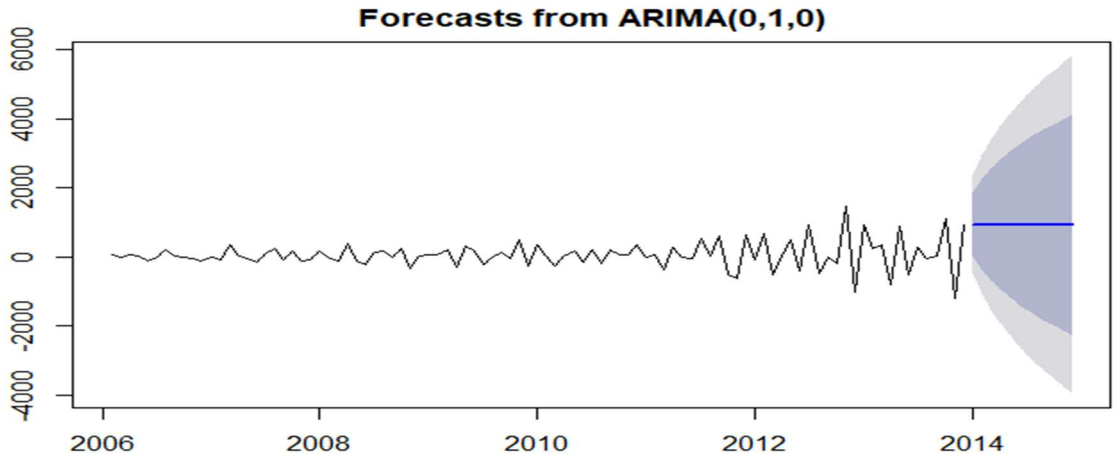


The Accuracy of the model is as follows

```
                     ME    RMSE      MAE       MPE     MAPE      MASE        ACF1
Training set 58.99465 278.66 193.4344 0.5779274 12.43197 0.2930242 -0.04949622
[1] "AIC =  1553.13594100915"
```

ARIMA Model: Autoregressive Integrated Moving Average (ARIMA) models include an explicit statistical model for the irregular component of a time series, that allows for non-zero autocorrelations in the irregular component.

Since ARIMA models are designed to work for stationary time series, Augmented Dickey-Fuller Test is performed on the time series and its established that the series is non-stationary. Series is made stationary by applying diff function. Once the series is made stationary auto arima model is applied



The accuracy of the model is

```
                     ME     RMSE      MAE MPE MAPE      MASE       ACF1
Training set 9.032358 717.5434 481.3481 NaN  Inf 0.9894753 -0.7350926
[1] "AIC =  1506.01184720415"
```

Though AIC is slightly better for ARIMA model, RMSE, MAE and MASE point to Holt-Winters model to be better forecasting model.

Month	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
Jan-14	5918.193	5526.991	6309.396	5319.901	6516.486
Feb-14	6196.302	5804.594	6588.009	5597.236	6795.367
Mar-14	6247.71	5854.867	6640.552	5646.909	6848.511
Apr-14	6299.653	5904.8	6694.505	5695.778	6903.528
May-14	6633.648	6235.676	7031.621	6025.002	7242.295
Jun-14	6543.377	6140.955	6945.8	5927.925	7158.83
Jul-14	6909.257	6500.855	7317.659	6284.66	7533.854
Aug-14	7008.554	6592.472	7424.637	6372.211	7644.897
Sep-14	7234.985	6809.38	7660.589	6584.079	7885.89

Oct-14	7333.343	6896.269	7770.416	6664.896	8001.789
Nov-14	7531.223	7080.662	7981.784	6842.149	8220.297
Dec-14	7563.77	7097.666	8029.875	6850.926	8276.615

CONCLUSION

The main outcome from the pilot project is that Fargo Health Group's Abbeville HC Incoming Examination future demand can be predicted using the Holts Winters Exponential Smoothing model. Also, it is established missing values in the data for this model can be imputed using MICE pmm method.

Also, it is recommended to perform this exercise for all the HCs as each of these HCs has a impact on another HC by means of rerouted requests. Though partial implementation may have positive impact on the HC for which this exercise is done, it may negatively impact the neighboring HCs by pulling resource from those HCs.

Data Quality – During the course of this pilot, some of the data issues identified points to the fact that there is no single data dictionary or data infrastructure across the enterprise. For example, each HC collects data in their own format and this pose a considerable issue when data is collected and merged from different HCs. Fargo Health Group needs to undertake a separate exercise to standardize the data collection process across the HCs.

Ethical Issues – It is also a need to be note that predicting and allocating health care professionals to the HC comes with an ethical dilemma with it. Consider the scenario in a long run the model detect one of the HC to keep losing its demand and ultimately may lead to closure of the center thereby affecting all the disabled persons who use to visit these centers for their tests. In another scenario where the model predicts a high demand for a HC thereby deploying more resources to this HC from neighboring HC. And when the prediction fails in this case, we end up pulling Health Care professionals from other centers but not utilize them.