Business Forecast of Fargo Health Group

Sanzida Parvin

Summary

The Fargo Health Group, started from 1961, is one of the most reputed health care service provider in the United States. This report represents the business problem and the importantance of data analysis to solve the problem related to disability compensation, a very special service of Fargo group. Upon receiving the disability test request from the patients, Health Centers (HCs) of Fargo, conduct the examination and provide the results within mandated 30-day timeframe. However, with increasing number of test requests over the years, HCs often do not have the capacity to meet the timeline with their resource limitations. Therefore, they sometime send some of the test requests to the neighboring out-of-network Outpatient Clinics (OCs), which costs around $1250 more than that of the in-house examination expenditure. Moreover, there were no guarantee that the OCs will finish the examination with in the required time frame. This clearly represents a huge financial and reputational risk for the Fargo group. The objective of this study was to find out the solution to mitigate the potential risk with the appropriate data analytical approach. The historical data analysis should help in precise prediction of number of upcoming test request and better scheduling of the examining physicians. The data analysis process comprised of three steps: (1) Data cleaning and detecting the missing values, (2) Imputed the missing values for analysis, and (3) Forecast the next 12 months data to predict the future volume of the request. Data cleaning process was performed by Microsoft Excel and imputation and forecast of the data was done by R programming language. Three forecasting methods, namely, Simple Exponential Smoothing, Double Exponential Smoothing and ARIMA have been applied to predict the numbers of upcoming test requests. Based on the accuracy and performance, ARIMA gives the best forecasting model. Therefore, it can be concluded that the reputational and financial risks of the Fargo group will be mitigated through efficient scheduling of the examining physicians based on the predicted upcoming test requests from ARIMA model.

Business Problems and Necessity for Data Analysis

The Quality Assessment Office (QAO) of Fargo is responsible for the collection of disability examination data from the 34 clinics and the subsequent analysis of that information. The disability examination process starts with the patient submitting a request for disability compensation to one of the organization's 34 Local Offices (LOs), and it is mandated by Fargo management that the Health Centers (HCs) complete and send back the results of disability examinations within 30 days after the receipt of the request from the LO. In actuality, however, due to the lack of examining physicians, HCs often do not have the capacity to meet the mandated 30-day timeframe. In such circumstances, the HC sometimes returns the request to the requesting LO right off the bat, together with the explanation that the rejection stems from the HC's being understaffed. Then the LO then reroutes the request either to other Fargo HCs in the vicinity (an infrequent scenario) or, more frequently, to one of the neighboring out-of-network Outpatient Clinics (OCs) with the hope that the OC will find the available staff and perform examinations on a timely basis. But it costs on average $1,250 more than the amount that Fargo would pay for an in-house examination of the request if there had been adequate in-house capacity. In addition, there were no guarantee that the OCs will finish the examination with in the required time frame as OCs are not in Fargo's network. Thus, such rerouting of requests from HCs to OCs represents yet another major financial and reputational burden for the organization.

In order to mitigate of this potential financial and reputational risk, Fargo's QAO Director, Jay Rubin, stressed the importance of an accurate, data-driven planning of examining physicians at the HCs. According to the Mr. Rubin, analyzing the historical data can provide clues on what may be happening in the future which could potentially lead to a more effective scheduling of examining physicians at the HCs and lessen the reputational and financial damages.

Data Analytic Approaches

The data analysis process for Fargo Health Group has included data cleaning, finding and imputing the missing values for analysis, and finally forecasting the next 12 months data. This analysis can help to predict the future work load and, hence, scheduling the examining physicians to ensure that the incoming disability examination requests will be performed by the in-house physicians within the required time frame. Microsoft excel has been used for the data cleaning process and R programming language has been implied to impute and forecast the data.

Nature and Structure of the Received Data, and the Perspective Solutions

The data provided for the analysis were unorganized, had some missing values as well as the wrong inputs. A part of the information related to Health Centers were disorderly and inaccurately distributed among the columns in the excel file. Moreover, the date of the collected test requests was not maintained in a proper way.

To resolve the problems associated with the provided data, as a first step, the missing and the unusual values has been detected and cleaned using filter method of MS Excel. Then some missing values were imputed by analyzing the given instructions and data sheets. For this purpose, filter method and conditional formatting of MS Excel has been applied. A cleaned version of the expected data set were obtained from the given data set, yet, some values are missing for the execution of analysis.

To impute the missing values R programming language has been used. The package of the R language named MICE (Multivariate Imputation via Chained Equations) was implemented in the data imputation process. MICE is one of the commonly used package by R users for creating multiple imputations as compared to a single imputation. It is popular for taking care of uncertainty in multiple missing values of time series analysis. As the data set was a time series variable, so MICE is one of the best package in R for imputing multiple missing values of a time series model. Before imputing data through MICE, a time series graphical presentation of cleaned data has been viewed to see the pattern of the data. Figure1 shows the time series plot of the cleaned data with the missing values.

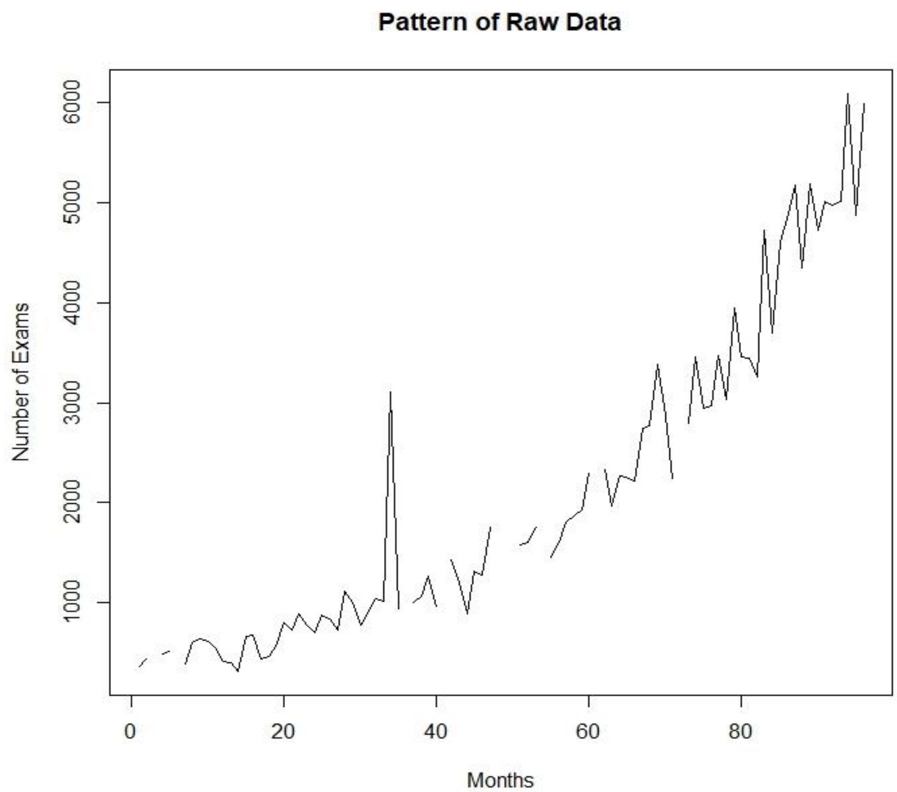

Figure1: Time series plot of the cleaned data with missing values

Form the figure, it is shown that the slop of the time series has an upward trend but not continuous, indicates the increment of the number of examination with the months goes on. The analysis shows a unusually large increment of the number of examination around the month 34, which was the consequence of closing the neighboring HC due to the natural disaster.

In MICE package, single or multiple data, can be imputed at a time using statistical equations. The R syntax for imputing the missing values by MICE is,

imputed_data = complete(mice(abbeville_data))
imputed_data

A graphical presentation (Figure 2) of the full data set has been performed in order to compare the trend of the time series of data before and after the imputation.

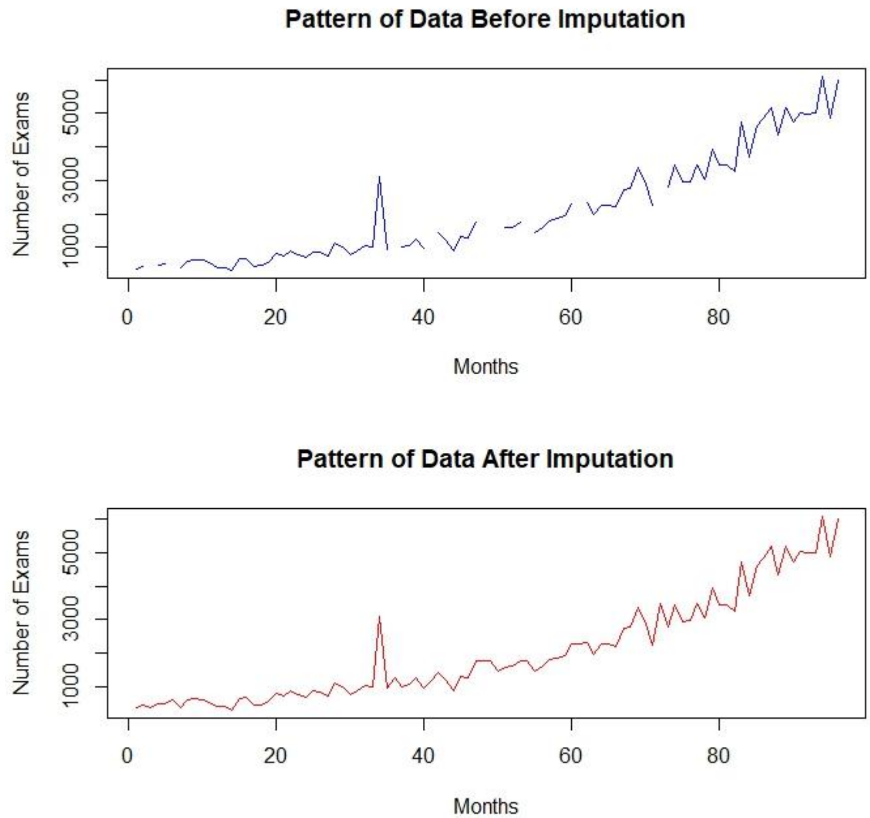

Figure 2: Time series plot of data before imputation and after imputation

Figure 2 clearly shows that the data patterns before and after the imputation are pretty similar, demonstrating the successful data imputation process.

Data Analysis and Forecasting:

Different analytical processes of data forecasting for the next 12 months has been compared to get the better values, so that the Fargo Health Group can mitigate their reputational and financial risk by proper scheduling of the examining physicians. The variables for this data analysis method were the number of requests for the disability test that were made for each of the month and the consecutive number of months.

A careful observations is required to select the data forecasting method as all the methods does not provide the significant values. As it is a time series data for this analysis, therefore, three time-series forecasting methods has been evaluated, and the best two methods has been determined on the basis of accuracy. These three forecasting methods are;

1. **Simple Exponential Smoothing**

   The outputs and accuracy of simple exponential smoothing are given below

```
> fit_data
ETS(A,N,N)

Call:
 ets(y = imputed_data$Request, model = "ANN")
  Smoothing parameters:
    alpha = 0.4161

  Initial states:
    l = 427.571

  sigma:  433.856

     AIC     AICc      BIC
1610.138 1610.399 1617.831

> forecast_value <- forecast(fit_data, 12)
> forecast_value
    Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
 97       5529.373 4973.364 6085.382 4679.031 6379.715
 98       5529.373 4927.150 6131.597 4608.352 6450.395
 99       5529.373 4884.237 6174.509 4542.723 6516.024
100       5529.373 4844.006 6214.740 4481.195 6577.552
101       5529.373 4806.009 6252.737 4423.084 6635.663
102       5529.373 4769.911 6288.835 4367.876 6690.870
```

103	5529.373	4735.452	6323.294	4315.176	6743.570
104	5529.373	4702.428	6356.318	4264.670	6794.076
105	5529.373	4670.674	6388.073	4216.106	6842.641
106	5529.373	4640.052	6418.695	4169.274	6889.473
107	5529.373	4610.450	6448.296	4124.002	6934.745
108	5529.373	4581.772	6476.974	4080.143	6978.604

Lo 80, Hi 80, Lo 95 and Hi 95 represent the 80% and 95% lower and higher confidence limit, respectively.

> accuracy(forecast_value)

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	127.7162	433.856	285.4065	2.055389	17.05165	0.8285799	-0.249502

The forecast values of simple exponential smoothing give 80% and 95% confidence intervals with AIC value of 1610.138 and the three forecasting errors, MAD = 285.4065, MAPE = 17.05165 and MASE = 0.882858

The graphical presentation of the forecasting values of simple exponential smoothing is given in Figure 3.

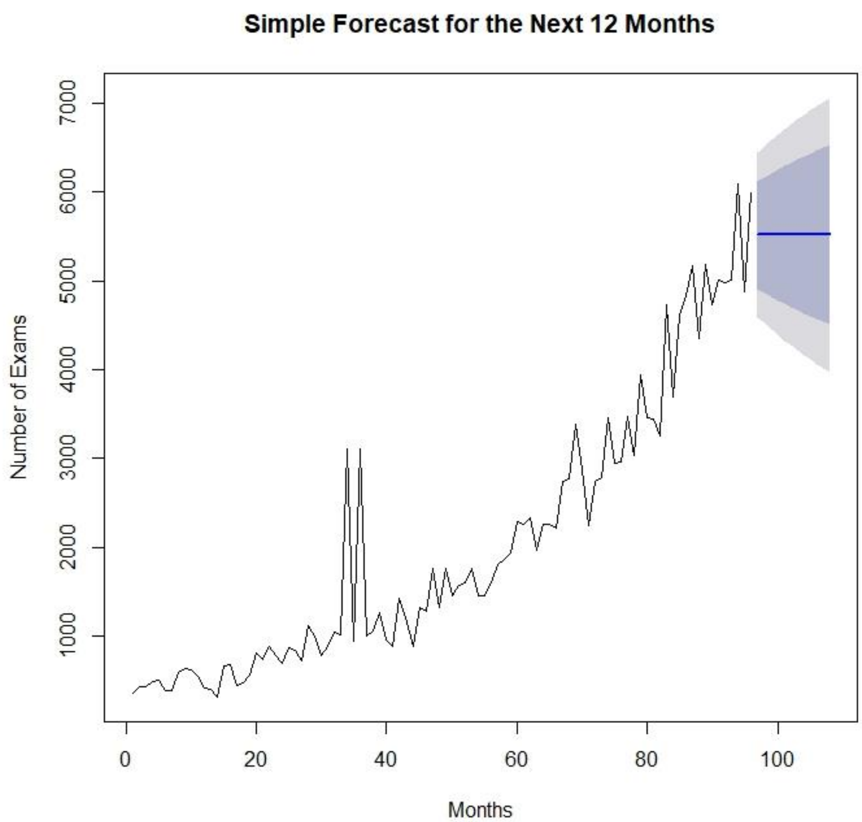

Figure 3: 12 months forecast of Abbeville data using simple exponential smoothing

2. Double Exponential Smoothing

The outputs and accuracy of the double exponential smoothing method of forecasting are given below

```
> fit_double
ETS(A,A,N)

Call:
 ets(y = imputed_data$Request, model = "AAN")

  Smoothing parameters:
    alpha = 0.1409
    beta  = 0.0222
```

Initial states:

l = 406.5142

b = 9.4432

sigma: 390.698

AIC	AICc	BIC
1594.021	1594.688	1606.843

```
> forecast_double <- forecast(fit_double, 12)
> forecast_double
```

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
97	5803.158	5302.458	6303.857	5037.404	6568.912
98	5934.524	5427.201	6441.847	5158.641	6710.408
99	6065.891	5550.142	6581.639	5277.121	6854.660
100	6197.257	5671.130	6723.384	5392.616	7001.898
101	6328.623	5790.050	6867.197	5504.946	7152.301
102	6459.990	5906.815	7013.164	5613.982	7305.997
103	6591.356	6021.375	7161.338	5719.645	7463.068
104	6722.723	6133.707	7311.738	5821.901	7623.544
105	6854.089	6243.817	7464.361	5920.758	7787.420
106	6985.455	6351.732	7619.179	6016.259	7954.652
107	7116.822	6457.499	7776.145	6108.474	8125.169
108	7248.188	6561.177	7935.200	6197.495	8298.881

Lo 80, Hi 80, Lo 95 and Hi 95 represent the 80% and 95% lower and higher confidence limit, respectively.

```
> accuracy(forecast_double)
```

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	57.10123	390.698	271.4224	-1.502172	16.94608	0.7972397	-0.1003454

The forecast values of double exponential smoothing have the smoothing parameter alpha = 0.1409, beta = 0.0222 with AIC = 1594.021 and the three forecasting errors, MAD = 271.4224, MAPE = 16.94608 and MASE = 0.7972 or 0.80%

The graphical presentation of the forecasting values of double exponential smoothing is given in Figure 4.

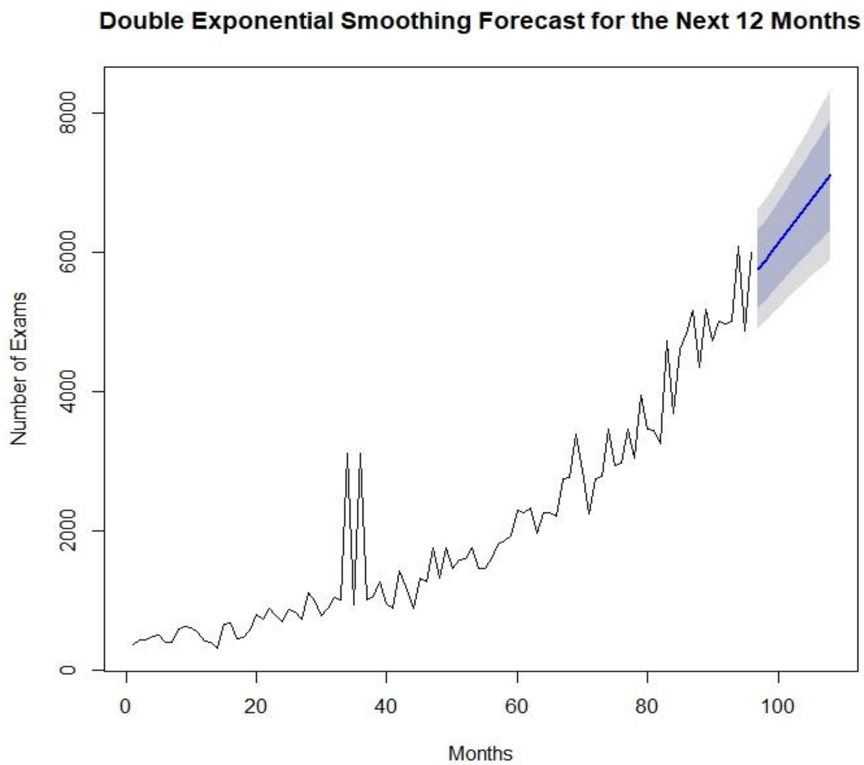

Figure 4: 12 months forecast of Abbeville data using double exponential smoothing

3. ARIMA Forecasting

Step by step outputs and accuracy from ARIMA forecasting method are shown below

Time series of the cleaned data with frequency measure of 1

```
> myTS <- ts(imputed_data$Request)
> myTS
> fit_arima <- auto.arima(x = myTS)
> fit_arima
```

Series:

ARIMA (1,1,1) with drift

Coefficients:

	ar1	ma1	drift
	-0.2506	-0.5855	54.7398
s.e.	0.1309	0.1043	13.9128

sigma^2 estimated as 166967: log likelihood=-704.87

AIC=1417.74 AICc=1418.18 BIC=1427.95

> forecast_arima <- forecast(fit_arima, h=12)

> forecast_arima

	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
97	5493.932	4970.269	6017.594	4693.059	6294.804
98	5687.479	5156.836	6218.121	4875.931	6499.026
99	5707.428	5141.906	6272.950	4842.536	6572.319
100	5770.888	5180.932	6360.844	4868.628	6673.147
101	5823.442	5208.100	6438.784	4882.358	6764.526
102	5878.729	5239.479	6517.980	4901.080	6856.379
103	5933.332	5270.922	6595.742	4920.263	6946.401
104	5988.106	5303.347	6672.866	4940.857	7035.356
105	6042.837	5336.429	6749.246	4962.478	7123.196
106	6097.579	5370.167	6824.992	4985.098	7210.061
107	6152.319	5404.492	6900.145	5008.616	7296.021
108	6207.059	5439.360	6974.757	5032.965	7381.152

Lo 80, Hi 80, Lo 95 and Hi 95 represent the 80% and 95% lower and higher confidence limit, respectively

> accuracy(forecast_arima)

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.49872	400.0127	266.5548	-10.00872	19.16732	0.7738504	0.0002223649

The forecast values of ARIMA method has the AIC = 1417.74 and the three forecasting errors, MAD = 266.5548, MAPE = 19.16732 and MASE = 0.773804 or 0.77%

The graphical presentation of the forecasting values of ARIMA is given in Figure 5.

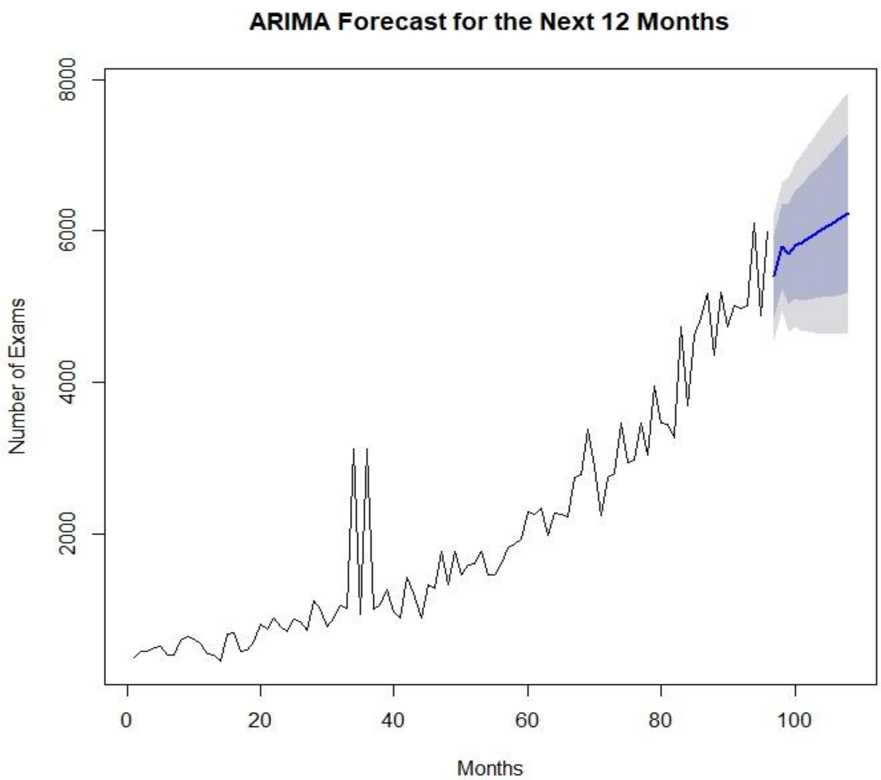

Figure 5: 12 months forecast of Abbeville data using ARIMA model

From the accuracy measurements of three forecasting methods, it can be concluded that the double exponential smoothing and the ARIMA model gives better forecasting than the simple exponential smoothing.

```
> accuracy(forecast_value)
                   ME    RMSE      MAE      MPE     MAPE      MASE      ACF1
Training set 127.7162 433.856 285.4065 2.055389 17.05165 0.8285799 -0.249502
> accuracy(forecast_double)
                   ME    RMSE      MAE       MPE     MAPE      MASE       ACF1
Training set 57.10123 390.698 271.4224 -1.502172 16.94608 0.7972397 -0.1003454
```

> accuracy(forecast_arima)

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	-0.49872	400.0127	266.5548	-10.00872	19.16732	0.7738504	0.0002223649

Also, the AIC values of the three methods suggest that the double exponential smoothing and ARIMA model fits better than the simple exponential smoothing method.

Simple exponential smoothing

AIC	AICc	BIC
1610.138	1610.399	1617.831

Double exponential smoothing

AIC	AICc	BIC
1594.021	1594.688	1606.843

ARIMA

AIC=1417.74 AICc=1418.18 BIC=1427.95

Ethical Implications

The objective of this data collection and analysis was to predict the upcoming test requests and, according to this prediction, to make a proper schedule for the examining physicians for completing the requests within the timeframe. By doing so, Fargo Health Group can regain their reputations and mitigate the future financial damages. The success of the new analysis report is depending on the proper and efficient uses of this model. The data for this analysis process was collected from the historical test requests. No personal information related to the patients was revealed for conducting this study. Therefore, in this particular study, the informed consent of patients might not be required before data collection. The sample used for this data analysis was the number of disability examination performed in last eight-year (96-months). Larger sample size usually gives better forecast. For doing this analysis, sample size of 96 is a reasonable number of sampling. This analysis was executed to forecast the upcoming number of disability examinations to make an efficient scheduling for the examining physicians to perform the tests within the required timeframe. So, this analysis can be used for all the parties who has the similar type of sample data and require similar type of forecasting. Fargo Health Group, who made this analysis by using their historical dataset was the owner of the dataset, analysis, and insights gleaned from data analysis. According to the analysis there were no moral obligations for Fargo Health to act based on the forecasting model. As the data analysis was performed by Fargo Health Group for their own purpose to give their patients a better service, so the Fargo Health Group is accountable for mistakes and unintended consequences in data collection and analysis.

Limitations

As being predictive analysis, it might have some limitations:

1. Missing values, even the lack of a section or a substantial part of the data, could limit its usability. For instance, the data might cover only one or two conditions of a larger set that are used for the modeling.
2. If patients condition or service providers condition changes unexpectedly, the methods might not be appropriate.
3. Time also plays a role in how well one technique works. Though a model may be successful at one point in time, but with time, customer's behavior changes and therefore a model must be updated.

References

1. https://hbr.org/product/fargo-health-group-managing-the-demand-for-medical-examinations-using-predictive-analytics/BAB266-PDF-ENG
2. https://uwli.courses.wisconsin.edu/d2l/le/content/3886059/viewContent/23748210/View
3. https://cran.r-project.org/web/packages/mice/mice.pdf