Shorter Waiting Time, Better Emergency Healthcare: Forecasting Stockholms' Emergency Department Visits

Martin Nordberg MD[1,2*], Oskar Eriksson B.Sc.[3]

1. Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet
2. Department of Surgery, Södersjukhuset
3. SAS Institute Inc., Sweden

ABSTRACT

Emergency department (ED) overcrowding is a problem for EDs world-wide, but first and foremost it is a problem for the patients who are affected from it as ED overcrowding leads to higher morbidity and mortality for the patients. Forecasting of ED inflow for planning ED operations and steering patients streams in a region is an integral part of mitigating the overcrowding issue. In the study at hand, four different forecasting models for ED inflow are developed, validated and tested. Gradient boosting as modelling method yielded the best results for forecasting the coming 72 hours of ED inflow, mean absolute percentage error of 21 %. This is a sufficiently accurate model to implement in a new information system for planning ED operations in a health care region.

Keywords: *Emergency department, overcrowding, forecasting, autoregressive integrated moving average, machine learning, gradient boosting, neural networks*

BACKGROUND

Emergency Department (ED) inflow is variable over time, but predictable to a large extent. To mitigate the problems arising from overcrowding in the ED, this variability is essential to understand (Eitel, Rudkin, Malvehy, Killeen, & Pines, 2010).

ED inflow, as defined as the number of visits to an ED over a time period, forecasting has been a subject of scientific studies for about 20 years(Holleman, Bowling, & Gathy, 1996; Rotstein et al., 1997). Forecasting the ED inflow allows for better planning of resources; human, economical and technical. This in turn, may, if used right, lead to less queuing in the ED and less overcrowding. Overcrowding has been shown to place patients at risk for poor outcomes and to cause prolonged pain and suffering for some patients (Derlet & Richards, 2000). A valid model for ED inflow forecasting is also useful in syndromic surveillance (Reis & Mandl, 2003). In addition to this, using forecasted ED inflow information may help Telephone Advice Services (TAS) to steer patients, geographically and temporally, between different ED.

A multitude of different mathematical models and a variety of variables have been used to forecast ED inflow for both longer and shorter amounts of time. The most commonly used model is AutoRegressive Integrated Moving Average (ARIMA) (Helfenstein, 1996). Amongst other models for ED inflow forecasting are linear regression and deep neural networks (Jiang, Chin, & Tsui, 2018; Wargon, Guidet, Hoang, & Hejblum, 2009). Using the machine learning (ML) method of gradient boosting as a forecasting method in other areas has shown promising results (Ayaru et al., 2015; Gala, Fernández, Díaz, & Dorronsoro, 2016; Zhang & Haghani, 2015). Forecasting ED inflow by using gradient boosting might be a way to gain a more accurate result than by previous, commonly used models.

A forecasting model with sufficient accuracy may be used both for planning operations and resources in the ED but also to steer patients through the TAS. In Stockholm County, a project for using the forecasting of ED inflow in the TAS is planned and will be commenced this spring.

AIMS

The aim of this study was to develop and evaluate methods of forecasting ED inflow for the coming 72 hours.

METHODS AND SETTING

SETTING

Stockholm County is the capital county of Sweden, with approximately 1,9 million inhabitants (Statistics Sweden.). The surroundings range from rural archipelago to urban inner city. The Stockholm County Council (SCC) is responsible for the health care system, which is universally free for all citizens Swedish and EU citizens.

The county has seven general emergency hospitals which range from 40,000 – 120,000 ED visits for adults per year. One hospital is a Trauma level I hospital.

DATA

Data on hourly presentation to the ED within the SCC was retrieved from the Value Analysis Layer (VAL) database which is a data warehouse for the SCC. Data in VAL is routinely collected from all different health care providers within the SCC, either public or private. The coverage ratio for VAL is within the 99^{th} percentile (Ekstrom, Kurland, Farrokhnia, Castren, & Nordberg, 2015; Wandell et al., 2013). VAL is used for follow-up and analysis of health care operations, for special projects like the project at hand and also for routine follow up conducted by the SCC itself.

Data on ED visit is produced whenever a patient is recorded in any of the different ED administrative data-systems for each hospital. Data is then transferred to VAL once every 24 hours. Among other administrative data such as age, sex and diagnosis for the visit, time stamps for when the visit started is recorded.

For the study at hand, three datasets were retrieved; training, validation and test sets.

The raw data contained a datetime variable containing the date and the hour of the day, and a variable with the count of actual visits for every hour in the data. From the datetime variable, a number of temporal variables were extracted. These were nominal/binary scaled variables indicating if the day in question is Christmas, New Year's Eve, Midsummer, or Valentine's day. There were also dummy-indicators for day of week, day of month, year, month, and hour of the day, as those variables, in previous research, have been shown to be significant for ED inflow (Hertzum, 2016)

One interval scaled variable is also created from the datetime variable; the distance from Christmas.

METHODS

From the count of actual visits variable, the 168 first lags were extracted, which equals a full week. The reasoning behind the week lag, was divided into two parts:

- The number of lags needs to be greater than or equal to the forecast horizon.
- They should also cover a relevant seasonality, and a lot of variation is to be found in the past weeks' behavior.

These were later input as either nominal or interval input variables in the predictive models.

Two different target variables were modeled: 1. the actual count which is interval scaled, along with 2. a binned variable of count, where 10 equally sized groups are built by ranking the actual count, this target variable is ordinal scaled, but since it was created with respect to group size it can be seen as approximately interval scaled. Technically, this is done through the RANK procedure in SAS (SAS Institute Inc.).

Four models were built using SAS/ETS 14.3 and SAS Visual Data Mining and Machine Learning (SAS Institute Inc., 100 SAS Campus Drive, Cary, NC 27513-2414, USA):

- ARIMA model with interval scaled target variable as a reference for the other models
- Gradient Boosting with interval scaled target variable
- Gradient Boosting with binned ordinal scaled target variable
- Neural Network with binned ordinal scaled target variable

The ARIMA model is described more thoroughly in the SAS documentation where also further references that describes the exact foundations of it are available (SAS Institute Inc.).

The gradient boosting used in SAS Viya is described in the SAS documentation (SAS Institute Inc.), with further links to the boosting method is based on, namely Hastie, Tibshirani and Friedman (2001) and Friedman (2001) (Friedman, 2001; Hastie, Tibshirani, & Friedman, 2001).

The neural network used in SAS Viya is described in the SAS documentation (SAS Institute Inc.) and is a multilayer perceptron neural network from Bishop (1995) (Bishop, 1995). The neural networks' architecture is a multilayer perceptron with direct connections between each input and each target neuron. Since the target is nominal, the target layer has a softmax activation function and an entropy error function.

All of these model types are set up with same partition for training and validation with an 85 % training and a 14.5 % validation split with the same seed for sampling. The last 0.5 % observations with known value for the target variable was used for testing the forecast performance, as in traditional time series forecasting. The tests are performed at the very end of the analysis in order to ensure an honest assessment of the forecast performance.

To optimize the models, they were all subject to auto tuning, meaning that a genetic algorithm sets out to find the architecture of the gradient boosting or the neural network that minimizes the information loss of the model. The auto tuning's only restriction was that it may not take more than 600 seconds in its search. SAS auto tuning is described in its documentation (SAS Institute Inc.)

The source data for the analyses consisted of 13848 observations, each observation being an hour. The starting hour is the 1st of January 2015 at midnight (i.e. the first hour of the year) and the last hour is the 30th of July 2016 at 23:00. The last 72 hours were removed and put into a separate test data set. 13776 observations remained for training and validation, split in each model as follows: The training set comprised 85% of observations and validation data comprised of 14.5 % of observations.

RESULTS

There was seasonality in ED inflow both for hourly, weekly and monthly variations. The inflow was highest during lunch time each day (Fig. 1), and the same was true for Mondays (Fig. 2). The summer, especially the vacation period showed a decreased daily inflow (Fig. 3 and 4).

Fig. 5 depicts the histogram of distribution of inflow numbers per hour.

ARIMA MODEL WITH INTERVAL SCALED TARGET

As a reference to be able to evaluate the machine learning models' performances, a simple ARIMA model is created. The model is fit with two autoregressive components, one of the first degree and another of the 24th degree, as well as a moving average component of the first degree. The model is fit to the first differences of the interval target to make it more stationary.

All parameters are highly significant as the conditional least squares estimation table (Table 2) show.

The residual correlation diagnostics (Figure 6) imply that the residuals are not a white noise process (i.e. completely random) and the residual normality diagnostics QQ-plot show deviations as well (Fig. 7). Together, these plots imply that the residuals are not likely to be independent and normally distributed.

Fig. 8 shows the forecast produced by the ARIMA procedure in correlation to the actual ED inflow.

To summarize model performance, summary statistics of the absolute deviations and the absolute percentage error is presented in Table 2.

GRADIENT BOOSTING WITH INTERVAL SCALED TARGET

The gradient boosting algorithm provides variable importance calculations as output. The top 5 variables from the gradient boosting were the 24^{th} lag, the 168^{th} lag, the first lag, the day of month and the 15^{th} lag.

Best configuration was 149 trees and 181 variables to try. The learning rate was 0.4135 and the sampling rate was 0.6991. The Lasso parameter was 8.5954 and the Ridge parameter was 5.7189. Mean squared error was 42.2931.

Since this model used an interval target it used mean squared error for optimization and tuning, it also yields an interval prediction, thereby making it difficult to compare with the other models. To remedy this, the predictions are put into bins following the same pattern as the previously binned count variable, creating comparable outputs.

The classification matrix, shows the relationship between the actual outcome and the binned forecast. A series plot comparing actual vs. binned forecast values better shows the relationship (Fig. 9).

To summarize model performance, summary statistics of the absolute deviations and the absolute percentage error is presented in Table 3.

GRADIENT BOOSTING WITH BINNED TARGET

The gradient boosting algorithm provides variable importance calculations as output. The top 5 variables from the gradient boosting were day of month, the 24^{th} lag of visit count, the first lag of visit count, the 168^{th} lag of visit count, and month.

Best configuration was 96 trees with 195 tested variables with a learning rate of 0.1436, a sampling rate of 0.4945, a Lasso of 0 and a Ridge value of 0.4293 and a misclassification error percentage of 7.38.

A classification matrix, shows the relationship between the actual outcome and the forecast. A series plot comparing actual vs. forecast values better shows the relationship (Fig. 10).

To summarize model performance, summary statistics of the absolute deviations and the absolute percentage error is presented in Table 3.

NEURAL NETWORK WITH BINNED TARGET

Best configuration: 2 hidden layers, 56 neurons in layer 1 and 78 in layer 2. L1 regularization parameter = 0 and L2 = 0.00004632 and the misclassification error percentage 44.83.

A classification matrix, shows the relationship between the actual outcome and the forecast. A series plot comparing actual vs. Forecast values better shows the relationship (Fig. 11).

To summarize model performance, summary statistics of the absolute deviations and the absolute percentage error is presented in Table 3.

DISCUSSION

In the present study, the number of visits in the near future where predicted with high accuracy. The gradient boosting model for interval data presented a Mean Absolute Percentage Error (MAPE) of 21 % when predicting the number of ED visits per hour for a 72-hour period. This result made gradient boosting out preform the other models in both this study and earlier research. Other forecasting models predicting the near future ED inflow on hourly basis reach an error (MAPE) roughly around 25 – 60 % (Hertzum, 2016; Jiang et al., 2018). As could be seen in Fig. 9 the model works best for the first 48 hours yielding a MAPE of 17.5 %. For clinical practice, knowing what to expect in the next 48 hours could be enough to redistribute resources and to free up hospital beds to meet the demand in the near future.

Using ARIMA to predict ED inflow yields a MAPE of 31,5 % which is in parity with earlier research. It strengthens the belief that gradient boosting could be a better model when forecasting ED inflow on other data sets.

Using ANN and gradient boosting on binned data to forecast ED inflow yield equivalent results. Gradient boosting is slightly more accurate with a MAPE of 27 % compared to ANN with a MAPE of 29,5 %. Both results are comparable with previous studies. Dividing the data in to bins before applying the GB-model did not show as good results as using the actual count of inflow. The reason for this might be due to the difference in scale. The binned data has an interval of 0 – 9 and the interval for the raw data was 0 – 232. This makes it hard to compare MAPE between the two models.

The results for longer forecasting periods as for a daily basis is much better and has been performed with a MAPE of 4 – 10 % (Hertzum, 2016, 2017; Jiang et al., 2018; Marcilio, Hajat, & Gouveia, 2013; Wargon et al., 2009). This was not tested in this study, so the accuracy of gradient boosting on ED inflow for longer time periods must be evaluated in future studies.

The temporal variation of ED inflow that was found in our data is well known and are consistent with earlier studies (Hertzum, 2016; Jones et al., 2009; Wargon et al., 2009). Calendar data was not among the top three variables for the Gradient boosting model. The most likely reason for this was that the calendar data effect was caught by the lag variables. Nothing in our dataset seem to deviate from the known and the expected.

ARIMA models are popular within time series analysis and forecasting. It is also taught at many universities as a “go-to” method. There are, however, prominent disadvantages with the ARIMA models along with the other time series models from traditional statistics (such as exponential smoothing models, econometric models, and regression models). These are 1) the models rely on distributional assumptions which need to be met and tested, and, in the case of

ARIMA specifically, 2) stationarity in terms of constant mean, variance, and covariance, which also needs to be fulfilled and tested. There are some complex strategies within econometrics to tackle these assumptions, but it is difficult. Furthermore, it is common knowledge that the more statistical hypothesis tests (of, for example, distributional assumptions) that are performed, the higher the probability of at least one of them being wrong. The need for assumptions are difficult to meet, through testing or other strategies, but most of all they make it almost impossible to put a trustworthy forecast into automated production.

In ML and more modern approaches to predictive modeling and forecasting, data is partitioned into training, validation and test data. The training data (which is normally the biggest part of the split data) is used to train models. To assess these models' performances, they are used on the validation data, meaning that if the trained model performs well on validation data, it *generalizes* well. Lastly, it is applied to test data which gives a final honest assessment of performance. The benefit from this is that distributional assumptions are not of interest anymore – if a model performs well on validation and test data, it is good for prediction. To do this, it is obvious that the modeler needs a lot of data points, and that is the new normal. When most time series analysis and econometric models were founded, the amounts of available data were normally lower and of higher quality. Statistics software were also less computationally efficient. As the ecosystem of data has moved from few data points and slow computers to vast amounts of data and fast computers, statistics must participate in that transition as well to stay relevant and value-adding.

The gradient boosting model in this study only used easily obtainable data to do the predictions. It also does not rely on any assumptions to perform. This makes it easy to automate and implement in clinical practice.

Using gradient boosting on ED inflow as a prediction model for hourly based predictions yields as good results that it may challenge ARIMA as the standard for inflow predictions in the future.

CONCLUSIONS

Using supervised ML methods for predicting ED inflow produce competitive results, and using classification methods seems to give even better results over commonly used non-ML methods. Gradient boosting may be favorable over neural networks for predicting the coming 72 hours of ED inflow to a health care-region.

REFERENCES

Ayaru, L., Ypsilantis, P. P., Nanapragasam, A., Choi, R. C., Thillanathan, A., Min-Ho, L., & Montana, G. (2015). Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting. *PLoS One, 10*(7), e0132485. doi:10.1371/journal.pone.0132485

Bishop, C. M. (1995). *Neural networks for Pattern Recognition*. Oxford: Oxford University Press.

Derlet, R. W., & Richards, J. R. (2000). Overcrowding in the nation's emergency departments: complex causes and disturbing effects. *Ann Emerg Med, 35*(1), 63-68.

Eitel, D. R., Rudkin, S. E., Malvehy, M. A., Killeen, J. P., & Pines, J. M. (2010). Improving service quality by understanding emergency department flow: a White Paper and position statement prepared for the American Academy of Emergency Medicine. *J Emerg Med, 38*(1), 70-79. doi:10.1016/j.jemermed.2008.03.038

Ekstrom, A., Kurland, L., Farrokhnia, N., Castren, M., & Nordberg, M. (2015). Forecasting emergency department visits using internet data. *Ann Emerg Med, 65*(4), 436-442.e431. doi:10.1016/j.annemergmed.2014.10.008

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Ann Stat, 29*(5), 1189-1232.

Gala, Y., Fernández, Á., Díaz, J., & Dorronsoro, J. R. (2016). Hybrid machine learning forecasting of solar radiation values. *Neurocomputing, 176*, 48-59. doi:10.1016/j.neucom.2015.02.078

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* . New York: Springer Verlag.

Helfenstein, U. (1996). Box-Jenkins modelling in medical research. *Stat Methods Med Res, 5*(1), 3-22. doi:10.1177/096228029600500102

Hertzum, M. (2016). Patterns in Emergency-Department Arrivals and Length of Stay: Input for Visualizations of Crowding. *The Ergonomics Open Journal, 9*(1), 1-14. doi:10.2174/1875934301609010001

Hertzum, M. (2017). Forecasting Hourly Patient Visits in the Emergency Department to Counteract Crowding. *The Ergonomics Open Journal, 10*(1), 1-13. doi:10.2174/1875934301710010001

Holleman, D. R., Jr., Bowling, R. L., & Gathy, C. (1996). Predicting daily visits to a walk-in clinic and emergency department using calendar and weather data. *J Gen Intern Med, 11*(4), 237-239.

Jiang, S., Chin, K. S., & Tsui, K. L. (2018). A universal deep learning approach for modeling the flow of patients under different severities. *Comput Methods Programs Biomed, 154*, 191-203. doi:10.1016/j.cmpb.2017.11.003

Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., & Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *J Biomed Inform, 42*(1), 123-139. doi:10.1016/j.jbi.2008.05.003

Marcilio, I., Hajat, S., & Gouveia, N. (2013). Forecasting daily emergency department visits using calendar variables and ambient temperature readings. *Acad Emerg Med, 20*(8), 769-777. doi:10.1111/acem.12182

Reis, B. Y., & Mandl, K. D. (2003). Time series modeling for syndromic surveillance. *BMC Med Inform Decis Mak, 3*, 2.

Rotstein, Z., Wilf-Miron, R., Lavi, B., Shahar, A., Gabbay, U., & Noy, S. (1997). The dynamics of patient visits to a public hospital ED: a statistical model. *Am J Emerg Med, 15*(6), 596-599.

SAS Institute Inc. The ARIMA Procedure. Retrieved from http://support.sas.com/documentation/cdl/en/etsug/63939/HTML/default/viewer.htm - arima_toc.htm

SAS Institute Inc. Base SAS(R) 9.2 Procedures Guide. Retrieved from http://support.sas.com/documentation/cdl/en/proc/61895/HTML/default/viewer.htm - a000146838.htm

SAS Institute Inc. The GRADBOOST Procedure. Retrieved from http://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=casml&docsetTarget=viyaml_gradboost_overview.htm&locale=sv

SAS Institute Inc. The NNET Procedure. Retrieved from (http://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=casml&docsetTarget=viyaml_nnet_overview.htm&locale=sv)

SAS Institute Inc. Shared Concepts. Retrieved from http://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.3&docsetId=casml&docsetTarget=viyaml_introcom_sect011.htm&locale=sv

Statistics Sweden. Folkmängd per distrikt, landskap, landsdel eller riket efter kön. År 2015 - 2016. Retrieved from http://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__BE__BE0101__BE0101A/FolkmangdDistrikt/?rxid=87519564-e99f-4194-9b98-62dc96678d64

Wandell, P., Carlsson, A. C., Wettermark, B., Lord, G., Cars, T., & Ljunggren, G. (2013). Most common diseases diagnosed in primary care in Stockholm, Sweden, in 2011. *Fam Pract, 30*(5), 506-513. doi:10.1093/fampra/cmt033

Wargon, M., Guidet, B., Hoang, T. D., & Hejblum, G. (2009). A systematic review of models for forecasting the number of emergency department visits. *Emerg Med J, 26*(6), 395-399. doi:10.1136/emj.2008.062380

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies, 58*, 308-324. doi:10.1016/j.trc.2015.02.019

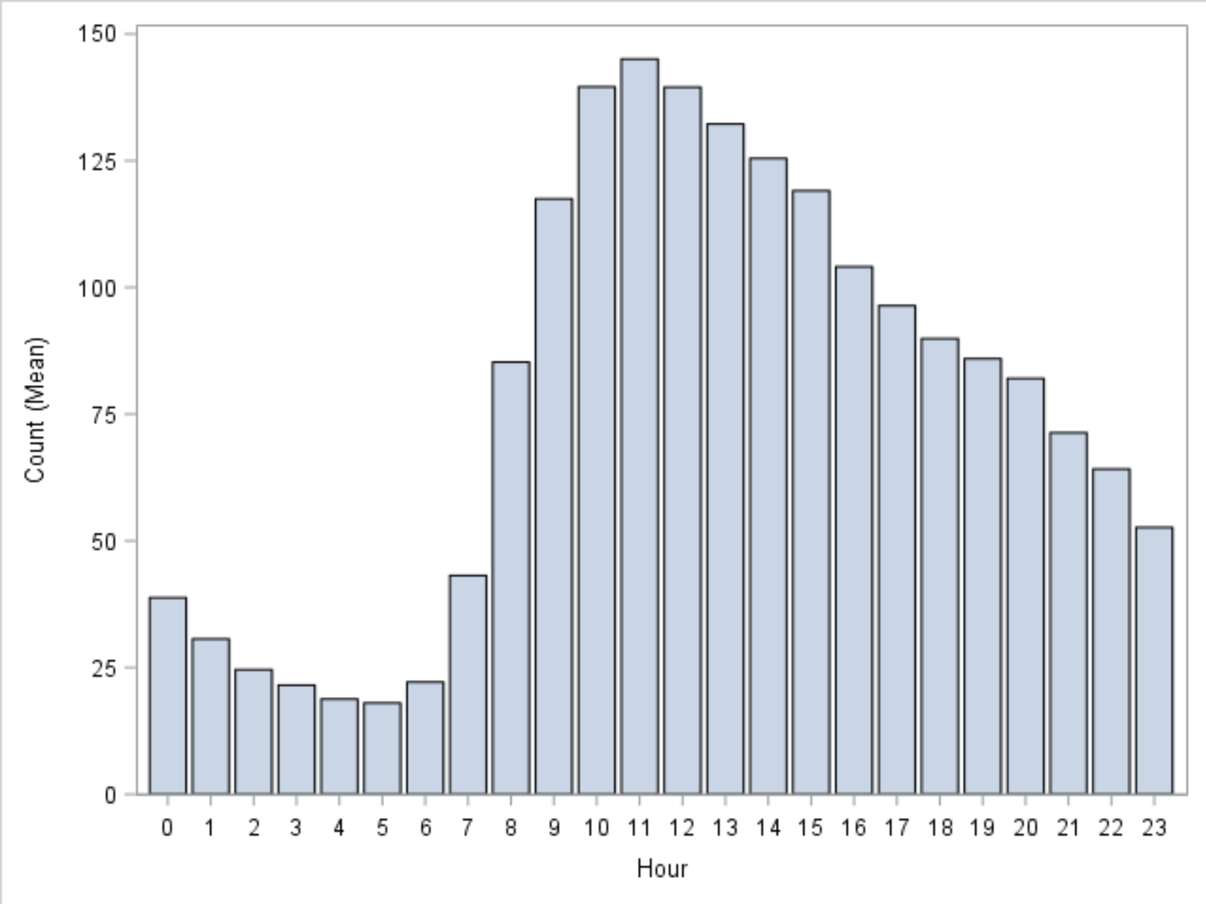

Figure 1. The mean ED inflow over hours of the day.

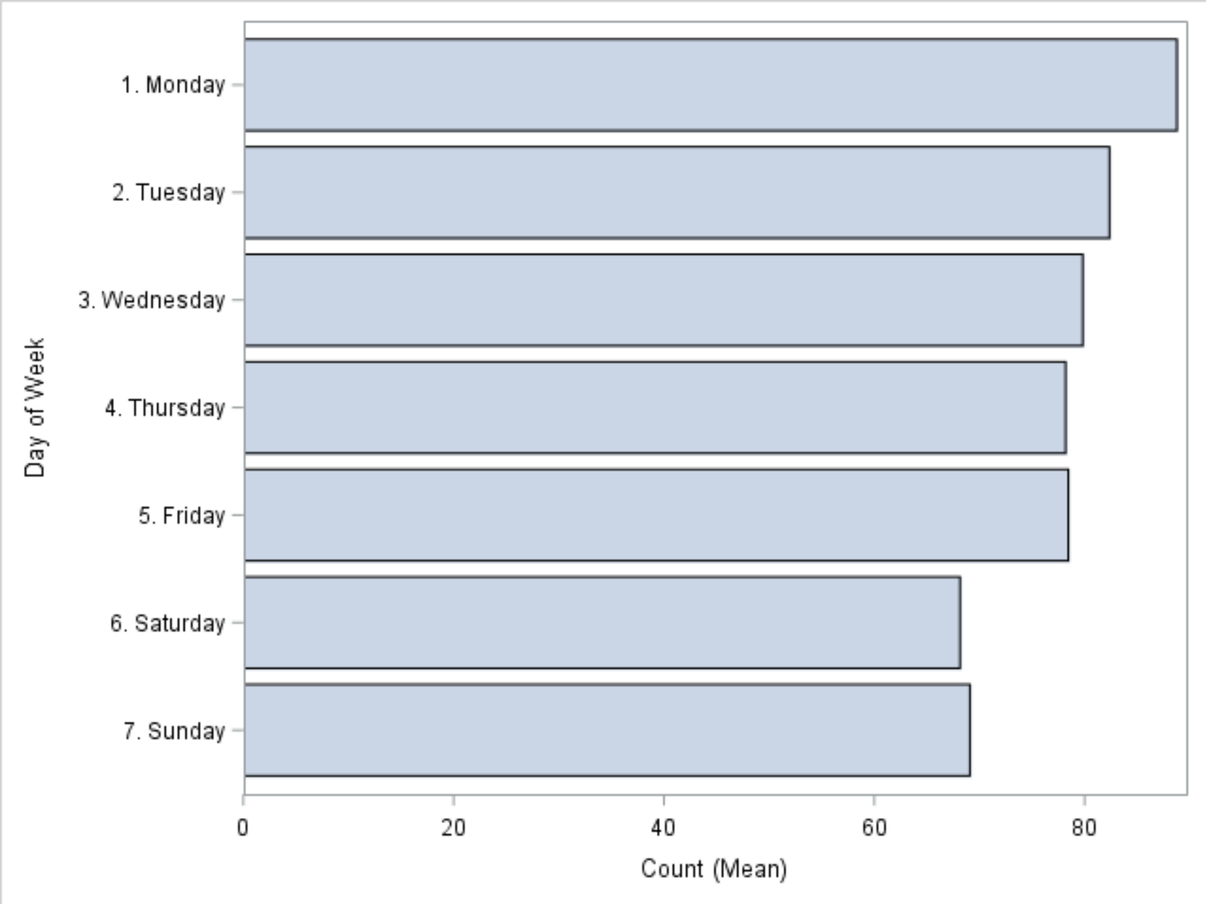

Figure 2. The mean ED inflow over the days of the week.

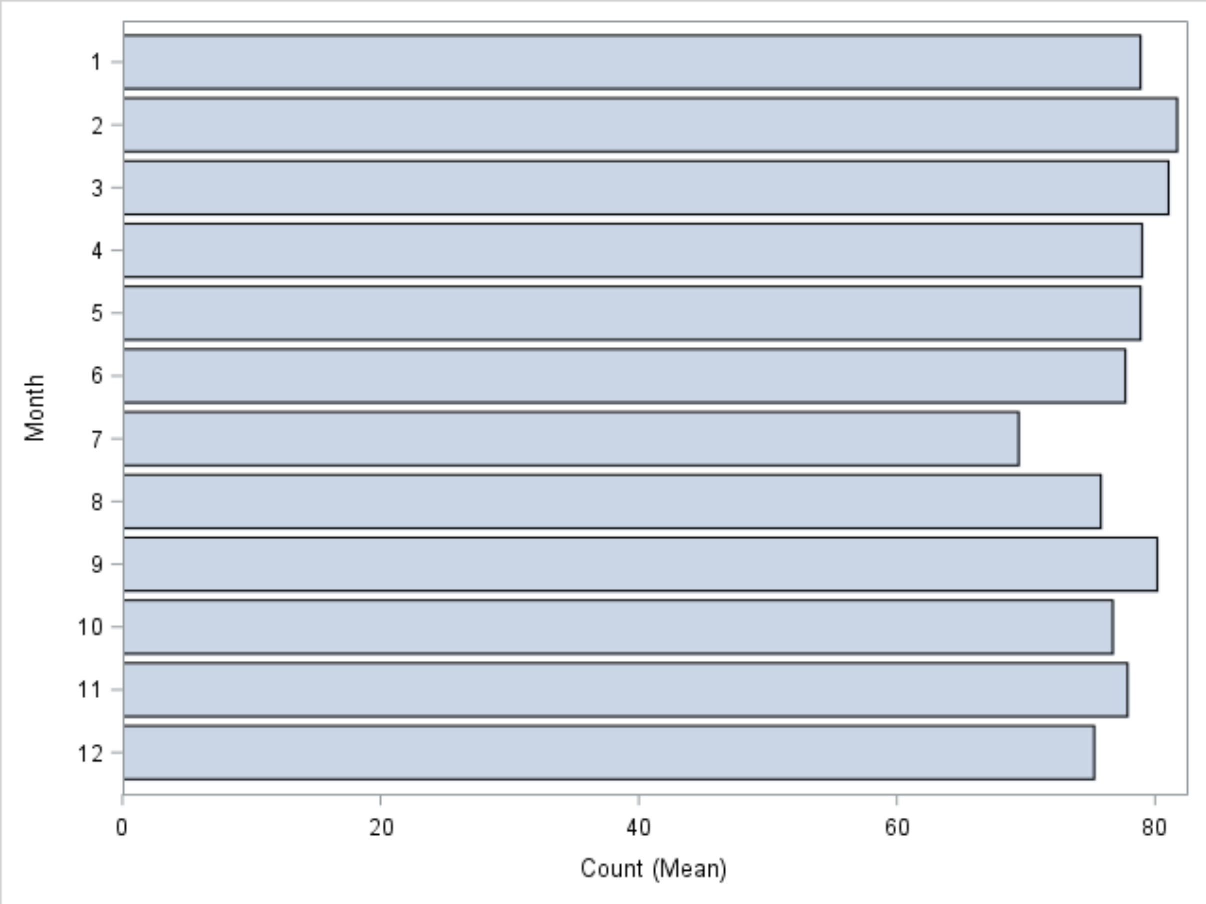

Figure 3. The mean ED inflow over the months of the year.

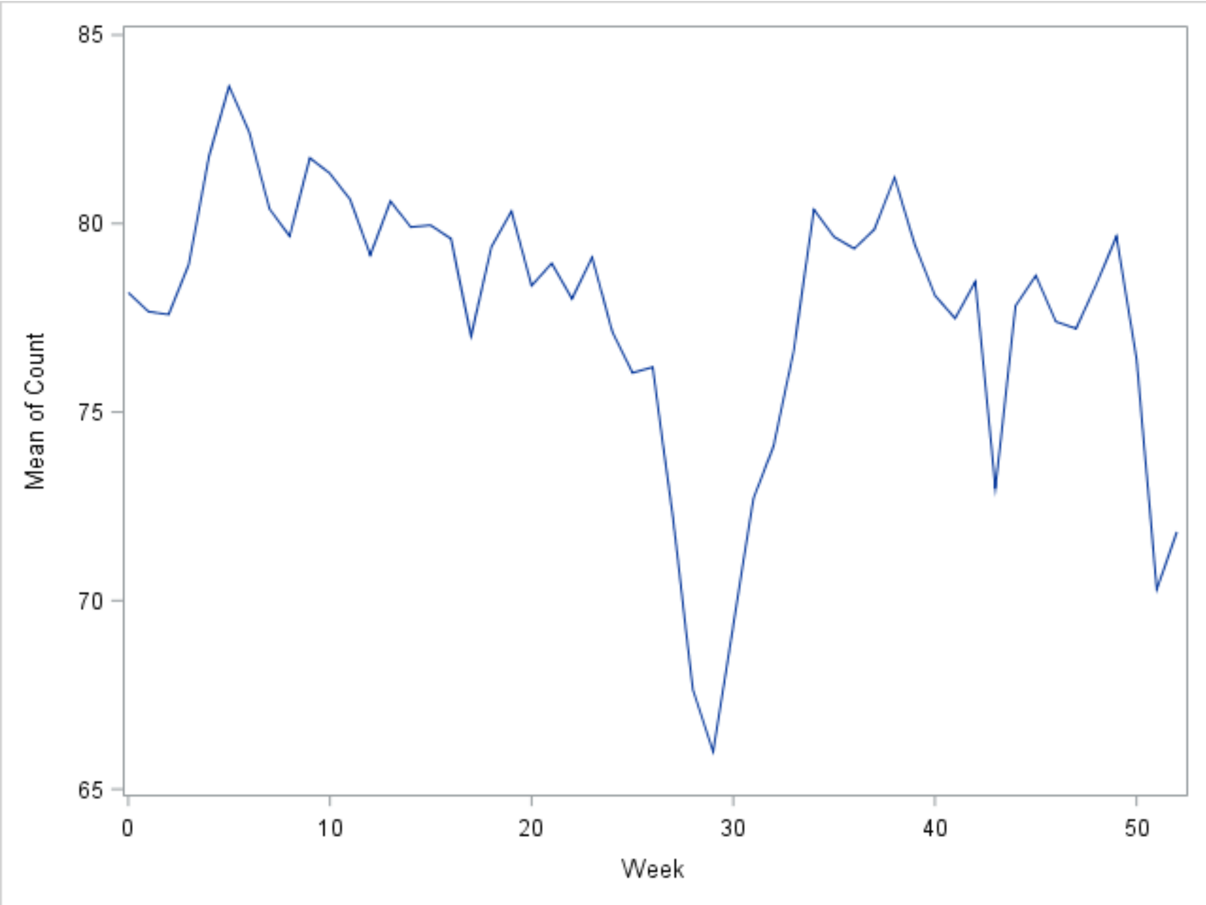

Figure 4. The mean ED inflow per week.

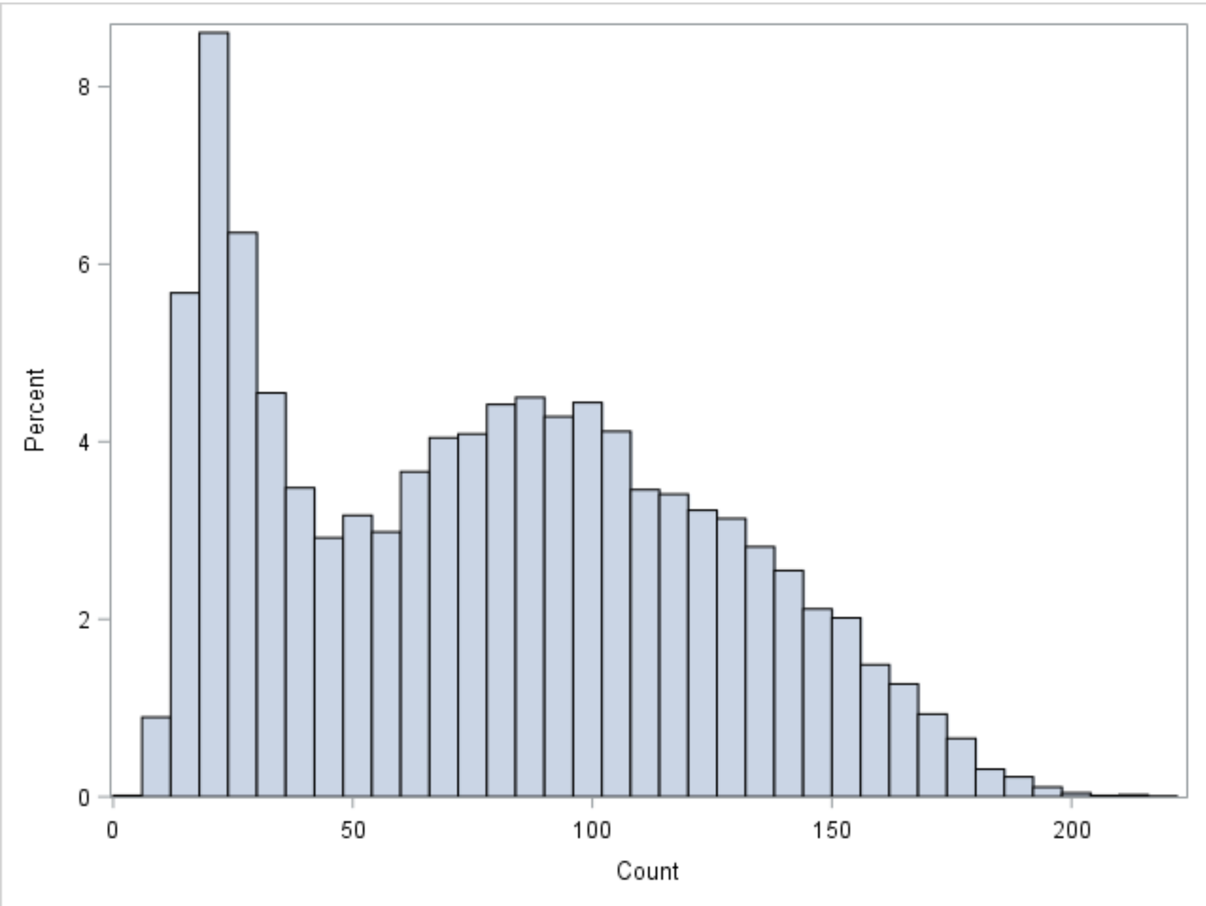

Figure 5. Histogram showing the distribution of number of visits to the EDs in Stockholm County per hour.

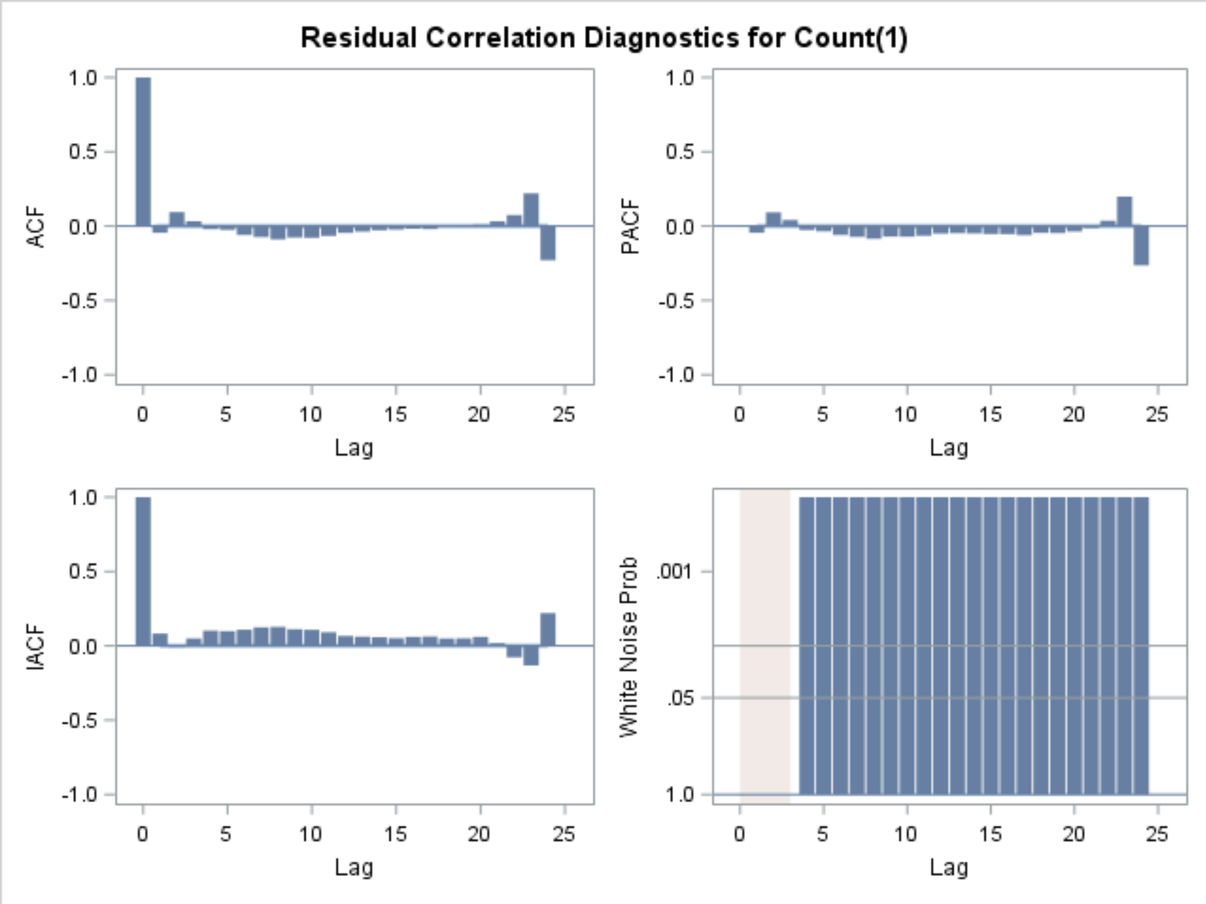

Figure 6. Residual Correlation Diagnostics for the ARIMA model

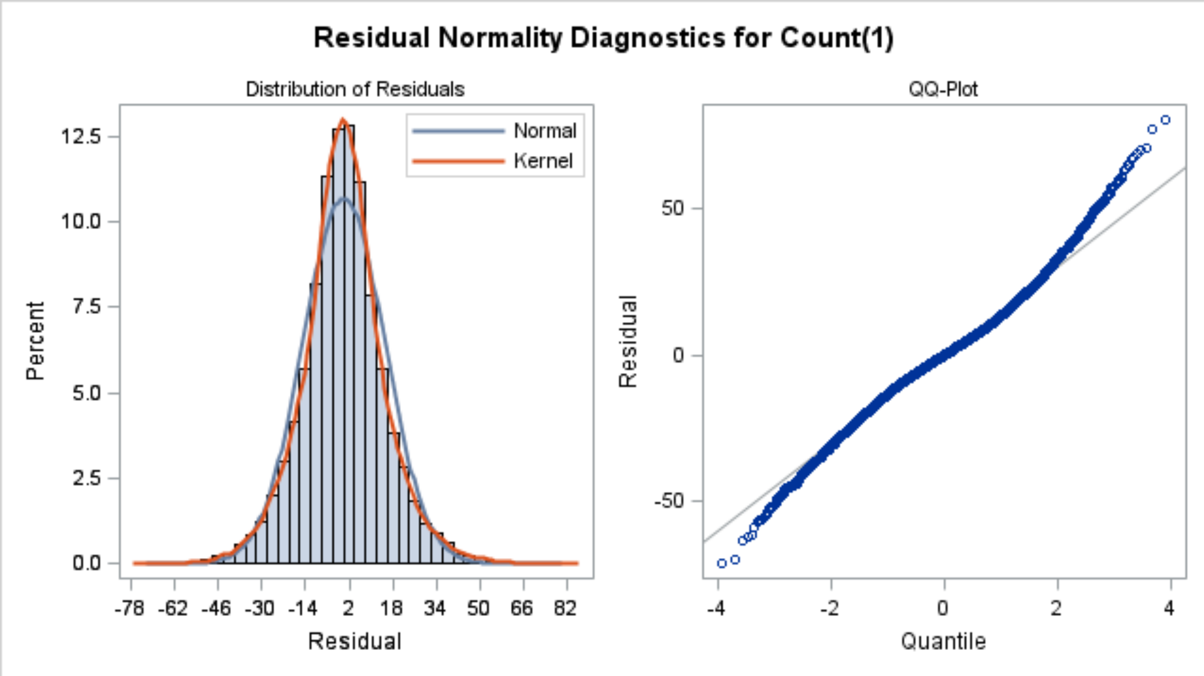

Figure 7. Residual Normality Diagnostic for the ARIMA model

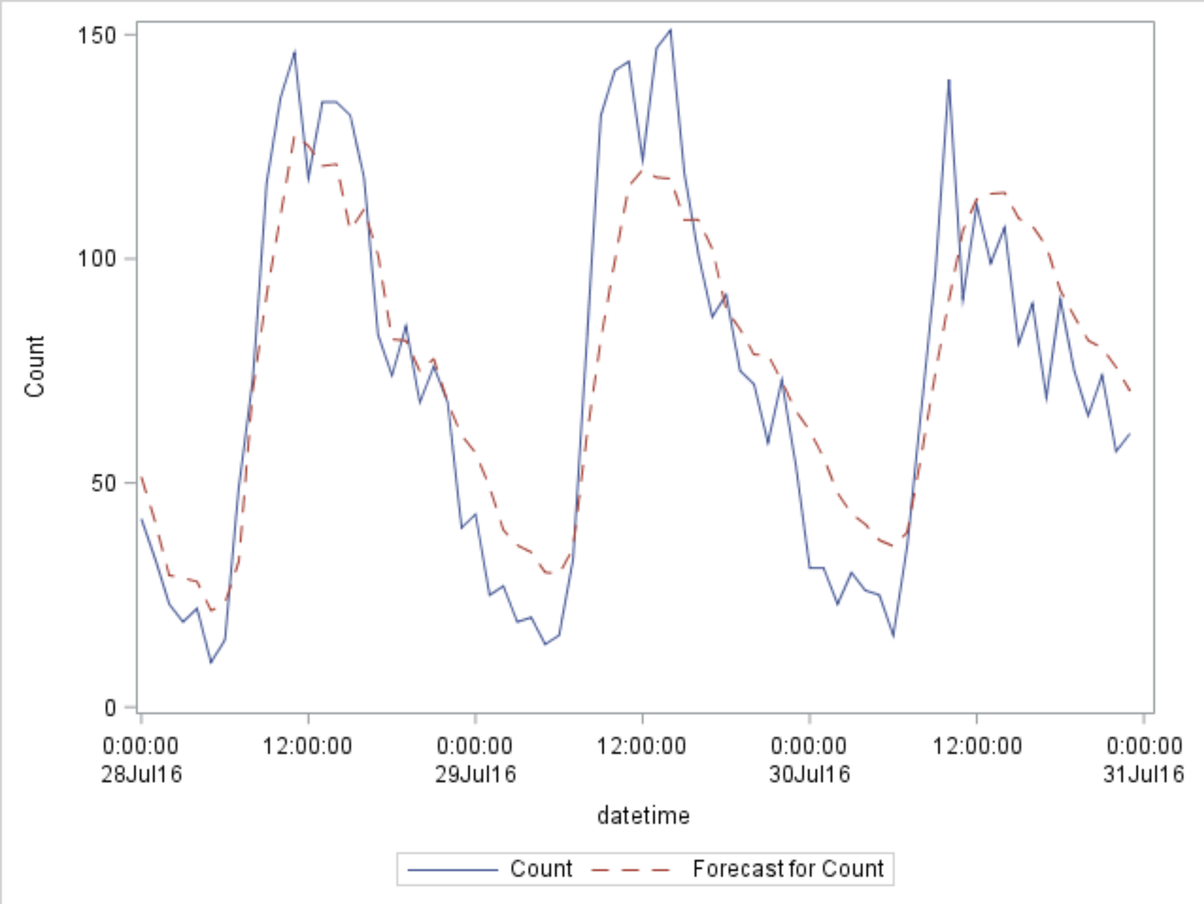

Figure 8. Actual vs. Forecasted count of ED inflow for the ARIMA model.

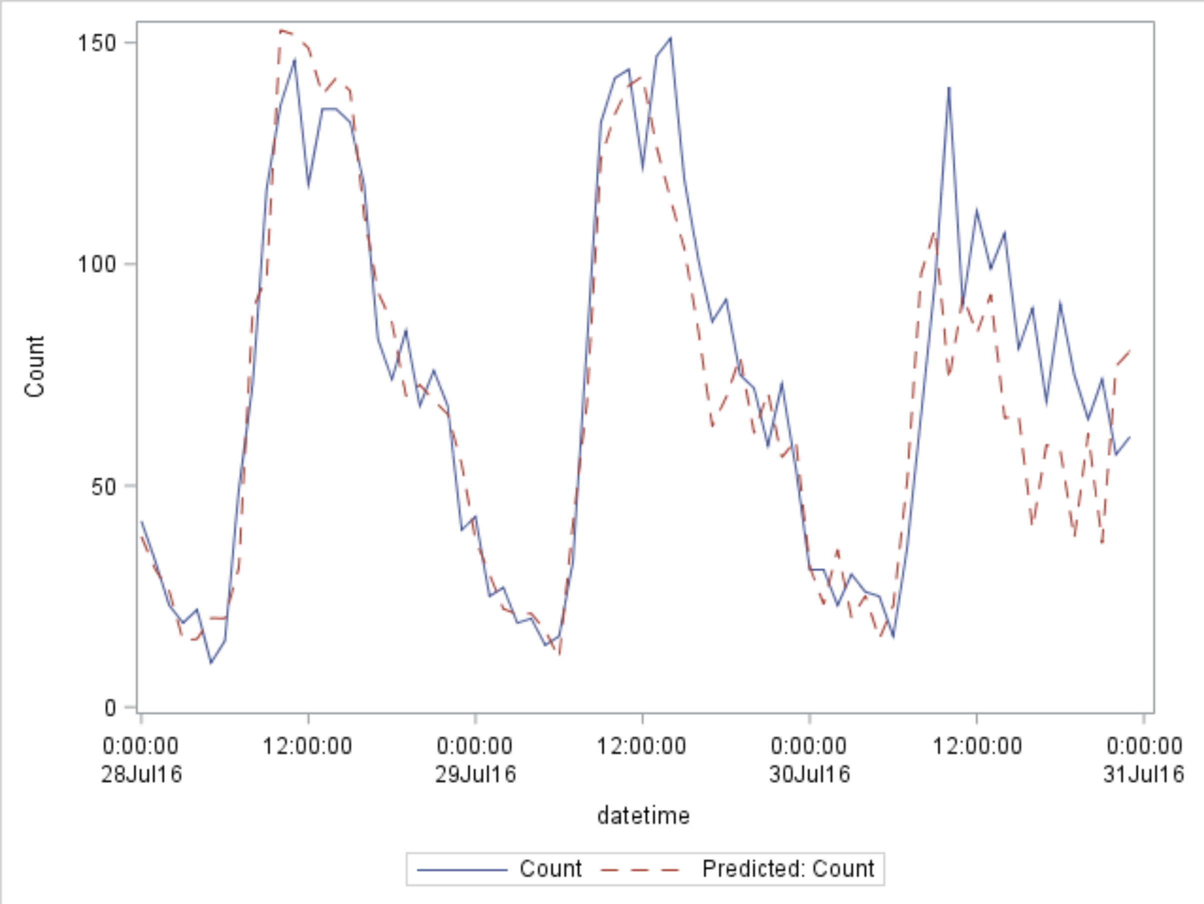

Figure 9. Actual vs. Forecasted count of ED inflow for the Gradient boosting model with interval scaled target.

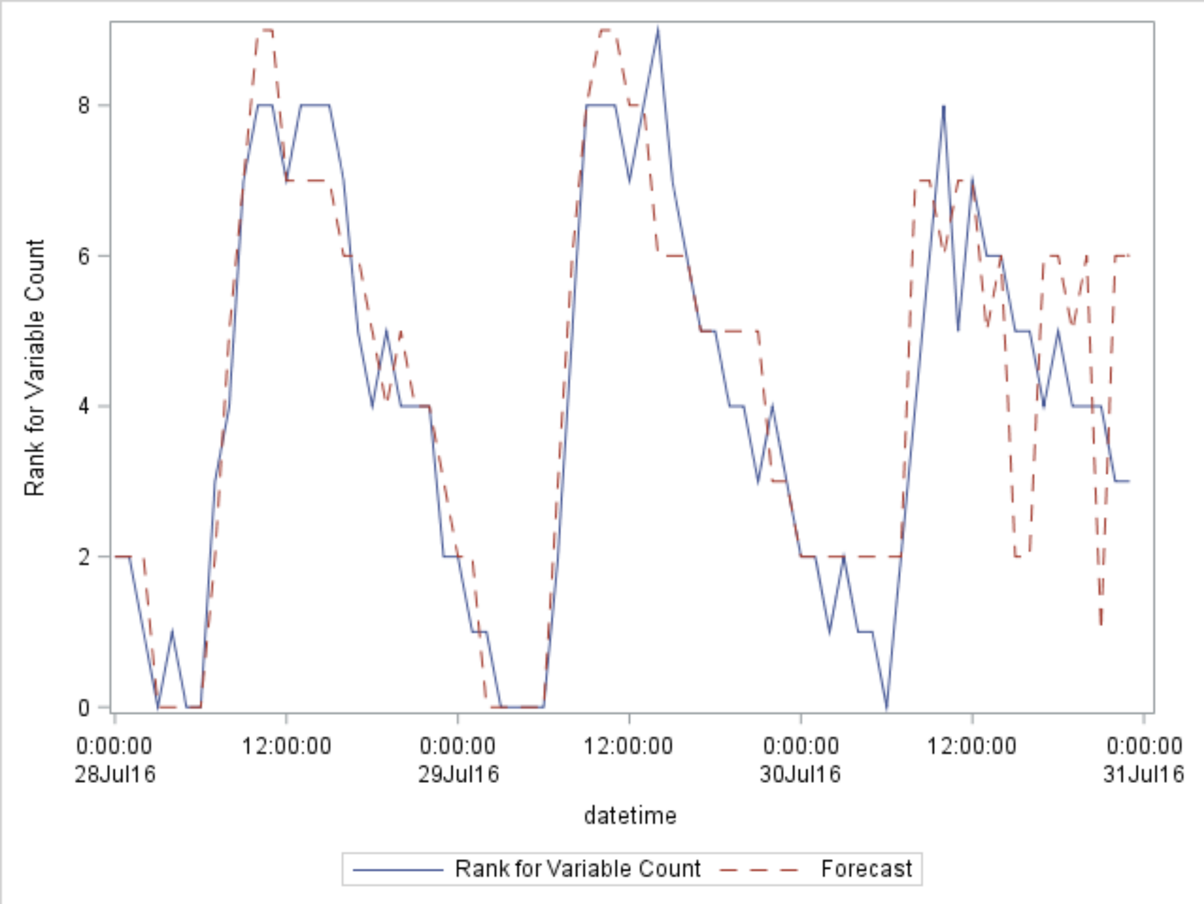

Figure 10. Actual vs. Forecasted count of ED inflow for the Gradient boosting model with binned target.

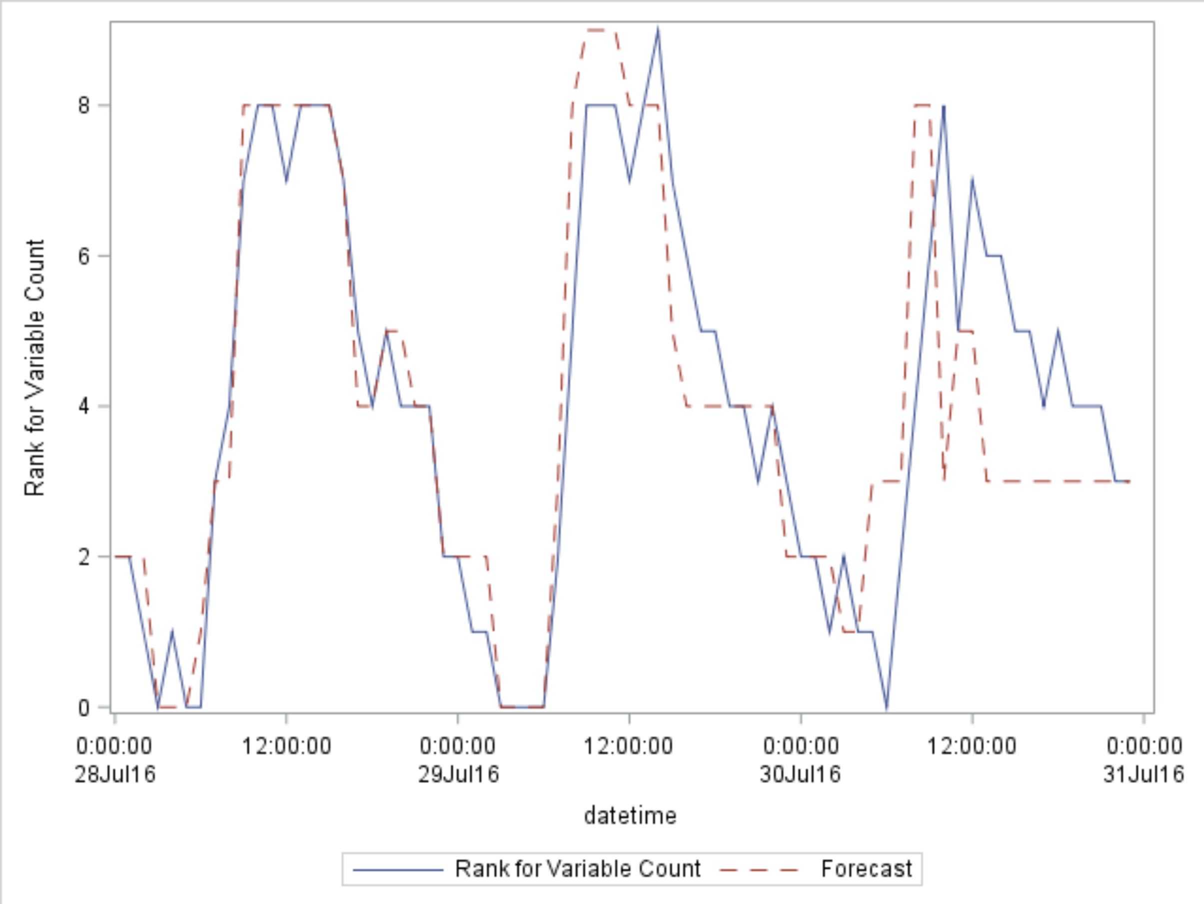

Figure 11. Actual vs. Forecasted count of ED inflow for the neural network model with binned target.

Bin	Number of patients per hour
0	0 - 20
1	21 - 28
2	29 - 43
3	44 - 63
4	64 - 79
5	80 - 94
6	95 - 108
7	109 - 126
8	127 - 147
9	148 - 232

Table 1. The size (ED inflow per hour) of the bins

Parameter	***Estimate***	***Standard Error***	***t Value***	***Approx Pr> |t|***	***Lag***
MU	-0.03	0.64	-0.05	0.96	0
MA1,1	0.63	0.01	71.45	<.0001	1
AR1,1	0.31	0.01	39.48	<.0001	1
AR1,2	0.62	0.01	93.61	<.0001	24

Table 2. Conditional Least Squares Estimation from the ARIMA model

Label	*Mean*	*Max*	*Min*	*Median*
GB interval target				
Absolute Deviation	13.60	66.30	0.64	9.83
Absolute Percentage Error	0.221	1.01	0.02	0.17
GB binned target				
Absolute Deviation	0.92	3.00	0	1.00
Absolute Percentage Error	0.31	1.00	0	0.20
NN binned target				
Absolute deviation	0.89	5.00	0	1.00
Absolute Percentage Error	0.28	2.00	0	0.17

Table 3. Summary Statistics for Absolute Deviation and Absolute Percentage Error Class Target. GB = Gradient boosting, NN = Neural network.