SESUG Paper 199-2023

Forecasting Goals for Student Success Metrics in Accountability Performance Reporting using SAS® Visual Analytics

Xiaoying Liu, Hennadii Balashov

University of South Florida

ABSTRACT

As a part of the State University System (SUS) of Florida, the University of South Florida is required to engage in the submission of an annual report, the SUS Accountability Plan, which monitors institution performance on a variety of metrics including student success, faculty research and awards, enrollment planning, and academic program development. It serves as an overall guidance for the university's five-year strategic plan. SAS® Visual Analytics Forecasting is utilized to project five-year goals on student success metrics of accountability plan based on the university's historical performance. By using the SAS® Visual Forecasting object, a time series forecasting model is generated for each of the 22 student success metrics. Visualizing data trends over time allows us to assess where the university stands at present, and where we want to go next. Forecasting analysis has been used to prepare the proposed goal rationale. Proposed goals are then provided to various campus stakeholders with analytics output to assist leadership in setting up university goals for accountability requirements. This paper will focus on the application of times series analysis in SAS® Visual Forecasting.

INTRODUCTION

The University of South Florida is a public research university, classified among "R1: Doctoral Universities – very high research activity". Recently, the university has accepted an invitation to join the Association of American Universities (AAU), achieving a monumental milestone in its development as one of the nation's leading research universities. As a part of the State University System of Florida, USF is required to submit an annual Accountability Plan to the Board of Governors, which reports the university's past five-year performance and next five-year's goals on metrics of students' success, faculty research and awards, enrollment planning and academic program development. Student success metrics track first-time-in-college students' fall-to-fall retention rate, academic progress rate (i.e., second fall retention rate with at least a 2.0 GPA), four-year and six-year graduation rates, average time-to-degree, transfer students' two-year and three-year graduation rates, university access rate (i.e., percent of undergraduates with Pell grant), as well as degree completion for minority students, and degree award in programs of strategic emphasis (USF Accountability Plan, 2023). Enrollment planning metrics keeps track of the university's headcount and FTE enrollment at different academic levels and by different student type categories, e.g. first-time vs. transfer students (USF Accountability Plan, 2023). The forecast is mainly focused on student success and enrollment planning metrics because we can access the university's historical data. We would like to analyze the institution's past performance, so we can confidently project where the institution is heading for the future. Time series analysis is used to forecast goals for the next five years.

TIME SERIES ANALYSIS

INTRODUCTION TO TIME SERIES ANALYSIS

Time series analysis is a powerful modeling technique used to predict future values based on historical data points, which are indexed in chronological order. The primary objective of this analysis is to discern trends and patterns within sequences of numerical data that are correlated with themselves but occur at different time intervals. This technique operates under the assumption that certain aspects of past patterns will persist into the future.

Time series forecasting holds wide-ranging applications across diverse industries, including weather forecasting, economic predictions, healthcare planning, engineering design, financial projections, retail inventory management, business planning, environmental research, and social studies analysis. If historical data is consistently timestamped, time series analysis methods can be employed for modeling, forecasting, and prediction. Concepts discussed in this section are primarily drawn from *Elements of Forecasting* textbook by Francis Diebold (Diebold, 1998).

MODELING APPROACHES IN TIME SERIES ANALYSIS

Time series analysis encompasses a multitude of modeling techniques, and selecting the appropriate one depends on the data characteristics. The two most widespread methods for forecasting are ARIMAX and Exponential Smoothing (ES) Models.

Autoregressive Integrated Moving Average with Explanatory Variable (ARIMAX) Model

The ARIMAX model can be likened to a multiple regression model enriched with autoregressive (AR) and/or moving average (MA) terms.

- The autoregressive (AR) component captures the lingering influence of previous data points, enabling us to forecast a variable's future values based on its past values. The term "autoregression" signifies that it is essentially a regression against itself. The essence of AR lies in expressing the current value as a linear combination of past values, with autoregressive coefficients determining the weights assigned to each past value. In simpler terms, knowledge of the series' past values enables predictions of future values. Below is an equation for a generic autoregressive process of order *p*.

  $$AR(p): \quad y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

  Where:

  - y_t: This is the value of the time series at time *t*. It represents the current value you are trying to model or predict.
  - $\phi_1, \phi_2, \dots, \phi_p$: These are the autoregressive coefficients. Each ϕ_i represents the weight or influence of the corresponding lagged value on the current value y_t. These coefficients determine how much the past values at different lags affect the current value. They are usually estimated from data.
  - $y_{t-1}, y_{t-2,} \dots, y_{t-p}$: These are the lagged values of the time series. These lagged values represent the historical values of the time series that are

used to predict the current value y_t.
  - ϵ_t: This term represents the error or residual at time *t*. It is the difference between the actual observed value at time *t* and the predicted value based on the autoregressive model.

- The 'I' in ARIMA refers to the process of integration, which eliminates non-stationarity in the mean function (i.e., trends) by differencing the data series. In time series analysis, a covariance stationary series is one in which the mean and variance remain constant over time, and the covariance between any two values depends solely on the time lag between them. Nonetheless, real-world time series often exhibit non-stationarity, attributed to trends or seasonality, causing shifts in mean and/or variance over time. To address non-stationary data, differencing is employed, involving the computation of differences between consecutive observations. This technique effectively removes trends or seasonality, rendering the series covariance stationary. In some cases, a single difference may prove insufficient, leading to the concept of an integrated process—a time series that has undergone differencing multiple times to achieve covariance stationarity.

- The moving average (MA) component considers the impact of past forecast errors, aiming to describe the autocorrelations within the data. The core idea behind MA is that the present value of the series results from a linear combination of its past prediction errors, with the coefficients determined by the moving average. In essence, knowledge of past prediction errors equips us to anticipate future values. Below is an equation for a generic moving average process of order *q*.

  $$MA(q): \quad y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}$$

  Where:

  - y_t: This is the value of the time series at time *t*. It represents the current value you are trying to model or predict, just like in the autoregressive process.
  - ϵ_t: This term represents the white noise or error at the current time step *t*. It is assumed to be a random variable with a mean of zero and constant variance.
  - $\theta_1, \theta_2, \dots, \theta_q$: These are the moving average coefficients. These coefficients determine how much the past error terms at different lags affect the current value. Like autoregressive coefficients, they are estimated from data.
  - $\epsilon_{t-1}, \epsilon_{t-2,} \dots, \epsilon_{t-q}$: These are the lagged error terms of the time series. These lagged error terms represent the historical errors in the time series that are used to model the current value y_t.
- The 'X' in ARIMAX denotes exogenous or independent variables, which can be incorporated to enhance the model's accuracy by considering external factors.

Application of the ARIMAX model involves a multi-step process of identifying, estimating, and validating a model that best fits a given time series data. While it can provide accurate predictions, it can also be quite tedious and time-consuming. The reason behind its tedious nature lies in the intricate steps required, including identifying the order of differencing, autoregressive, and moving average components, estimating model parameters, and performing diagnostic tests to ensure the model's validity. Additionally, fine-tuning and iterating on the model may be necessary, which can involve extensive trial and error. Fortunately, SAS Viya automates most of the procedures related to ARIMAX model implementation; more on this in the next section. ARIMAX model can also be implemented using SAS Enterprise Guide or SAS Studio; however, delving into the details of this implementation is outside the scope of the present paper. For comprehensive guidance on this topic, an excellent resource to consult is the SAS/ETS® 13.2 User's Guide, The ARIMA Procedure (SAS Institute Inc., 2014).

Exponential Smoothing (ES) Model

ES models calculate forecasts as weighted averages of past observations, with the weights diminishing exponentially as the observations age. In essence, more recent data points carry greater influence in these models. The basic equation for exponential smoothing is:

$$y_t = \alpha * x_t + (1 - \alpha) * y_{t-1}$$

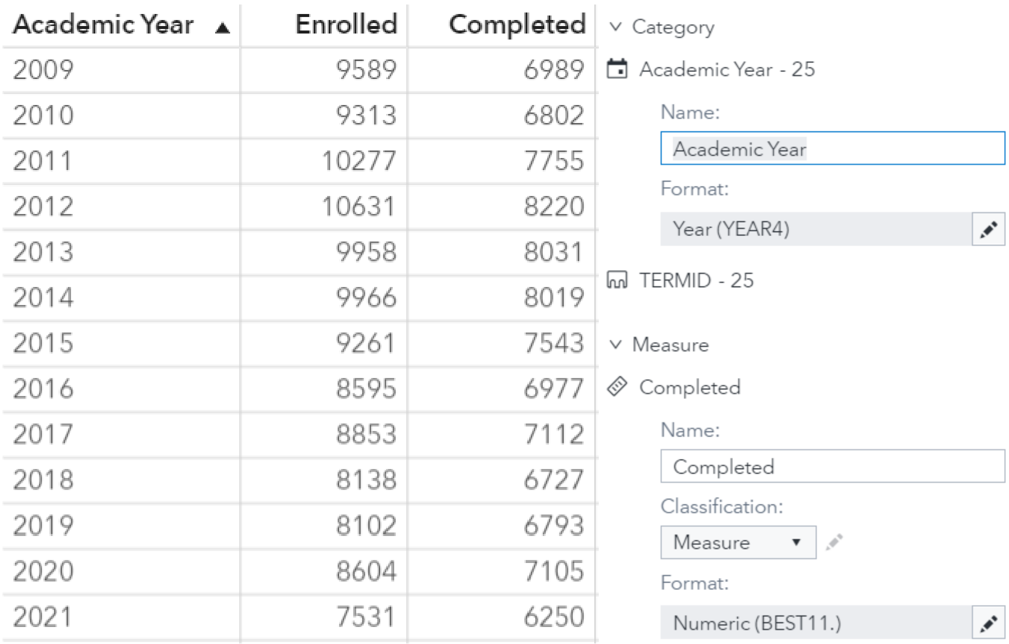
Where:

- y_t is the forecast for period *t* based on the actual observation x_t and the forecast for the previous period y_{t-1}.
- α is the smoothing parameter, which is a value between 0 and 1 that determines the weight given to the most recent observation x_t versus the forecast for the previous period y_{t-1}.

Higher α values prioritize recent data, yielding forecasts that are responsive to data fluctuations, while lower α values emphasize past forecasts, resulting in more stable and resistant predictions, less affected by short-term data variations.

APPLICATION IN SAS VISUAL ANALYTICS FORECASTING

The software we use is SAS® Visual Analytics, one of the modules within SAS® Viya applications. Forecasting object is used to generate time series analysis and forecast on the university's goals. In order to use the forecasting object, the first step is to restructure the data into a time series data set. The rows of the table must represent data over some period of time (for example, years or months) and the columns of the table contain at least one measure data item for forecasting.



Academic Year ▲	Enrolled	Completed
2009	9589	6989
2010	9313	6802
2011	10277	7755
2012	10631	8220
2013	9958	8031
2014	9966	8019
2015	9261	7543
2016	8595	6977
2017	8853	7112
2018	8138	6727
2019	8102	6793
2020	8604	7105
2021	7531	6250

Category

Academic Year - 25

Name: Academic Year

Format: Year (YEAR4)

TERMID - 25

Measure

Completed

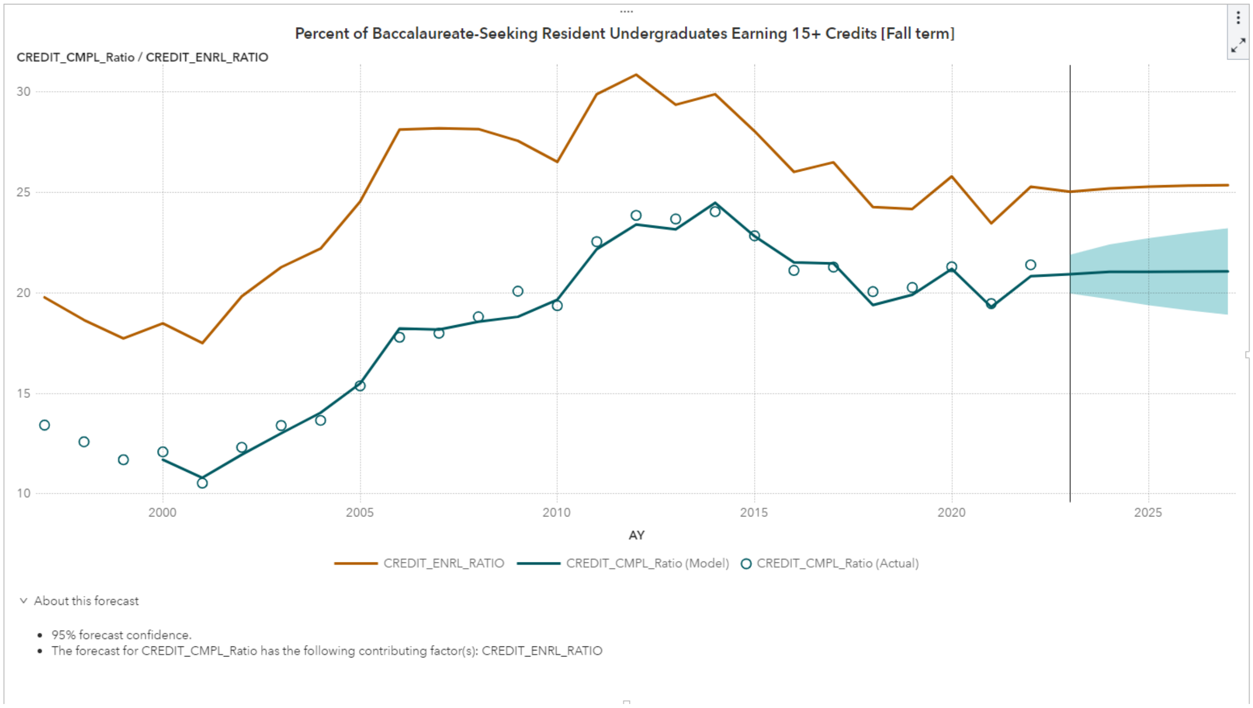
Name: Completed

Classification: Measure ▾

Format: Numeric (BEST11.)

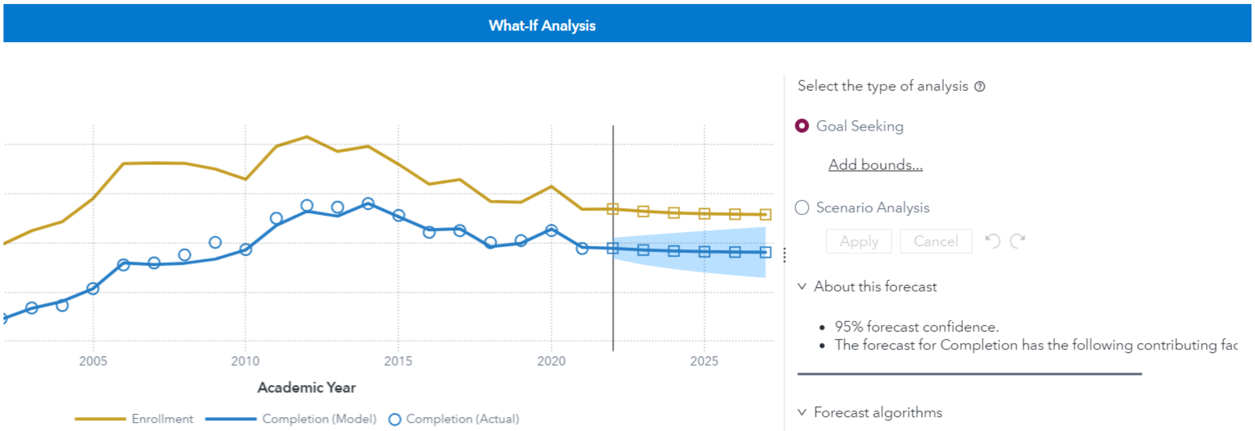
SAS® Visual Analytics automatically tests the forecasting models (i.e., ARIMA, ESM) based on the trend and pattern in the data, and selects the best model. The forecast displays a line with predicted values and a colored band that represents the confidence interval. By default, the next six periods are forecast, and the 95% confidence interval is displayed. Historical values for the forecasting model are displayed as markers only (without a line). Historical predicted values are displayed as part of the forecast line (Ball et al., 2020). For ARIMA model, if the input factor is significant, then the historical and future predicted value for the input factor will also be displayed in the model.

We add underlying or input factors in the model to narrow down the prediction interval. The underlying factor is the X component in ARIMAX model. These additional measures are evaluated by the model to determine whether they contribute to the accuracy of the forecast. If the additional measures do not increase the accuracy of the forecast, then they are not applied to the forecasting model. If the additional measures do improve the accuracy of the forecast, then the forecast line is adjusted, and the confidence bands are narrowed (Ball et al., 2020). For each metric, we add several underlying factors, and the model automatically picks the most significant ones. In the example below, the credit enrollment ratio (i.e., orange line) is found to impact the credit completion ratio (i.e., green line), and it is selected in the model as the contributing factor, and forecast is adjusted to include enrollment impact on completion.



- 95% forecast confidence.
- The forecast for CREDIT_CMPL_Ratio has the following contributing factor(s): CREDIT_ENRL_RATIO

With Forecasting object, we also do what-if analysis to manually adjust the prediction values. To perform What-If analysis for a forecast, measures need to be added to the underlying factors role and found to contribute to the forecast (Ball et al., 2020). There are two ways to do the what-if analysis. With scenario analysis, we could modify future values of the underlying factors to see what impact those changes have on the forecast. With goal seeking, we modify the future values of the forecast to see what changes in the underlying factors are needed to reach the goal. In both cases, we can either set future values equal to a certain number, or adjust series values by a constant or percentage. What-If analysis allows us to assess the possible outcomes due to certain events (e.g., policy change, pandemic) that might be favorable or unfavorable to the university’s performance. This is when we can incorporate some human thinking into the model building process to generate more realistic forecast.



Select the type of analysis

● Goal Seeking

Add bounds...

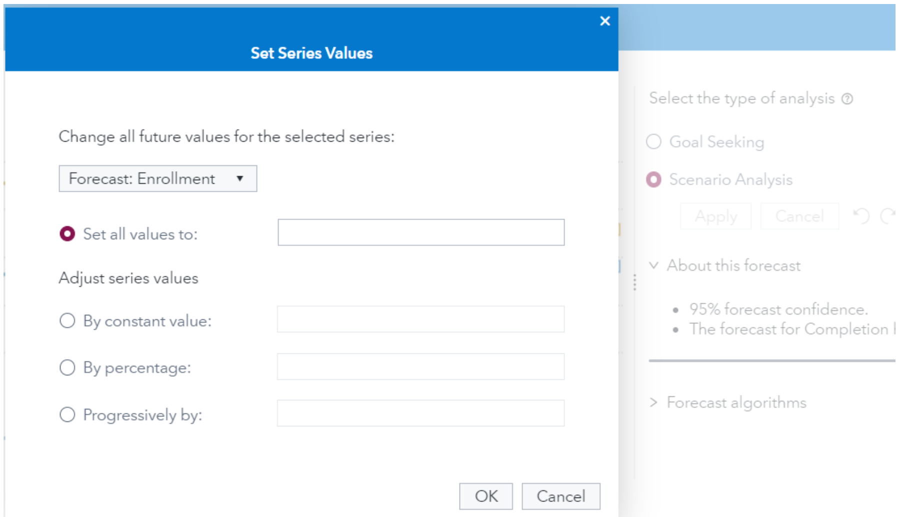
○ Scenario Analysis

About this forecast

- 95% forecast confidence.
- The forecast for Completion has the following contributing fac

Forecast algorithms

With scenario analysis, we modify the predictive values of the underlying factors.



Change all future values for the selected series:

Forecast: Enrollment ▾

● Set all values to:

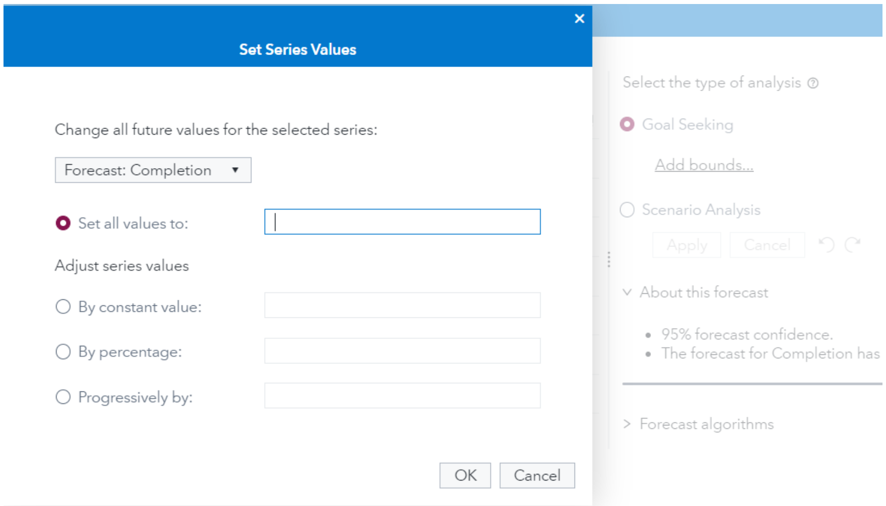
Adjust series values

○ By constant value:

○ By percentage:

○ Progressively by:

With goal seeking, we could set all predictive values of the forecast variable to a constant or adjust values by percentage.



Change all future values for the selected series:

Forecast: Completion ▾

● Set all values to:

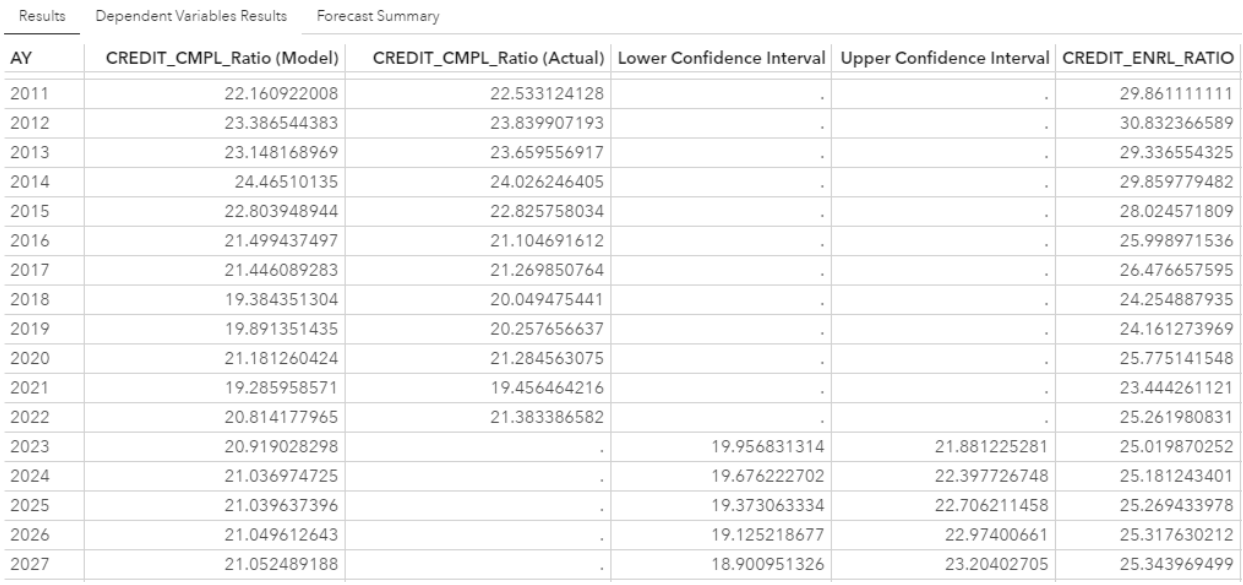
Adjust series values

○ By constant value:

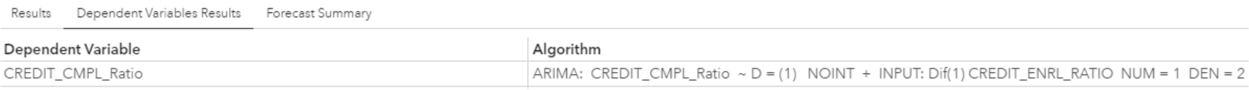
○ By percentage:

○ Progressively by:

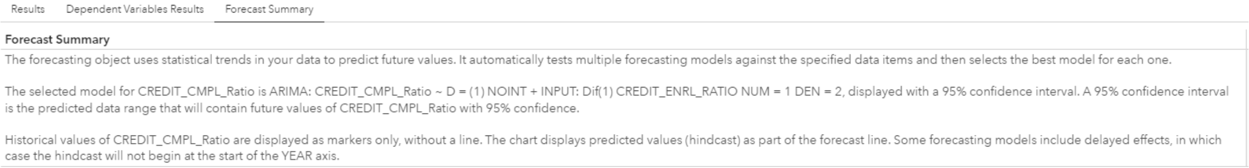
In addition to the graph, forecasting object also generates details about the forecast algorithm and summary. Historical predicted values and future predicted values are calculated using the forecasted algorithm selected for the data. For future values, a lower and upper confidence interval are also calculated. If underlying factor is included in the model, historical and predicted values will also be displayed in the table.



AY	CREDIT_CMPL_Ratio (Model)	CREDIT_CMPL_Ratio (Actual)	Lower Confidence Interval	Upper Confidence Interval	CREDIT_ENRL_RATIO
2011	22.160922008	22.533124128	.	.	29.861111111
2012	23.386544383	23.839907193	.	.	30.832366589
2013	23.148168969	23.659556917	.	.	29.336554325
2014	24.46510135	24.026246405	.	.	29.859779482
2015	22.803948944	22.825758034	.	.	28.024571809
2016	21.499437497	21.104691612	.	.	25.998971536
2017	21.446089283	21.269850764	.	.	26.476657595
2018	19.384351304	20.049475441	.	.	24.254887935
2019	19.891351435	20.257656637	.	.	24.161273969
2020	21.181260424	21.284563075	.	.	25.775141548
2021	19.285958571	19.456464216	.	.	23.444261121
2022	20.814177965	21.383386582	.	.	25.261980831
2023	20.919028298	.	19.956831314	21.881225281	25.019870252
2024	21.036974725	.	19.676222702	22.397726748	25.181243401
2025	21.039637396	.	19.373063334	22.706211458	25.269433978
2026	21.049612643	.	19.125218677	22.97400661	25.317630212
2027	21.052489188	.	18.900951326	23.20402705	25.343969499



Dependent Variable	Algorithm
CREDIT_CMPL_Ratio	ARIMA: CREDIT_CMPL_Ratio ~ D = (1) NOINT + INPUT: Dif(1) CREDIT_ENRL_RATIO NUM = 1 DEN = 2



Forecast Summary

The forecasting object uses statistical trends in your data to predict future values. It automatically tests multiple forecasting models against the specified data items and then selects the best model for each one.

The selected model for CREDIT_CMPL_Ratio is ARIMA: CREDIT_CMPL_Ratio ~ D = (1) NOINT + INPUT: Dif(1) CREDIT_ENRL_RATIO NUM = 1 DEN = 2, displayed with a 95% confidence interval. A 95% confidence interval is the predicted data range that will contain future values of CREDIT_CMPL_Ratio with 95% confidence.

Historical values of CREDIT_CMPL_Ratio are displayed as markers only, without a line. The chart displays predicted values (hindcast) as part of the forecast line. Some forecasting models include delayed effects, in which case the hindcast will not begin at the start of the YEAR axis.

For model validation, we look at the confidence interval and adjust model accordingly to narrow down the prediction interval. We also compare the model prediction with the actual numbers. Viya® gives us very reasonable estimates as you can see that the model prediction is very close to the historical numbers.

CONCLUSION

One of the most valuable applications using data visualization is seeing trends over time. One assumption underlying this forecasting technique is that some aspects of the past patterns will continue into the future. Time series analyses allow us to see where the institution stands at present, and where we want to go next. Forecasting results serve as the proposed goal rationale, and have been used to assist leadership to make goals for the university.

Adding input factors in forecasting allows us to better understand how different data elements relate to each other. In most cases, model selected significant input factors follow the similar pattern as the forecasting variable, for example, enrollment impacts completion. We also found that the percentage of first-generation students, and black/Hispanic students impact university access rate. Access rate is the percent of undergraduates with a Pell grant. Pell grant is a financial need-based federal grant, usually offers to low-income students. Many first-generation students and ethnic minority students come from low-income families. In order to improve access rate, the university may enhance recruitment strategies to attract more low income and ethnic minority students. Helping more low-income and minority students get education will boost local economic growth, benefit society through educating highly skilled graduates. Forecasting analyses not only provides evidence on the overall university's performance in pursuit of various goals to meet accountability requirements, but also allow us to see the potential areas for institutional growth.

REFERENCES

Ball, N., Bell, R., Hardin, B., Matthews, L., & Stemler, T. (2020). *SAS® Visual Analytics 2 for SAS® Viya®: Advanced Course Notes.* Cary, NC: SAS Institute Inc.

Diebold, F. X. (1998). *Elements of Forecasting.* Cincinnati, OH, USA: South-Western College Pub.

SAS Institute Inc. (2014). *SAS/ETS® 13.2 User's Guide.* Cary, NC: SAS Institute Inc.

Tabachnick, B. G., & Fidell, L. S. (2007). Using Multivariate Statistics (5th ed.). Boston: Allyn and Bacon

USF Accountability Plan (2023) https://www.usf.edu/ods/data-tools/accountability.aspx

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Xiaoying Liu, Ph.D.
University of South Florida
xiaoyingliu@usf.edu

Hennadii Balashov, M.S., M.A.
University of South Florida
balashov@usf.edu