# Statistics and Show Business: Shakespeare Meets Predictive Analytics

Xinping Zhang, Miami University, Department of Statistics
Byran J. Smucker, Miami University, Department of Statistics
Jay Woffington, Cincinnati Shakespeare Company

"There is nothing either good or bad but thinking makes it so."
-- William Shakespeare's *Hamlet*, Act II, Scene II

**1. Introduction**

A Board member, a marketing professional and an actor walk into a bar… The actor says, "Why aren't you spending more money on marketing my great Shakespearean performance?" The Board member concurs, "I agree. Why are our sales so slow? Surely, there's got to be something we can do." The marketing professional responds, "First, the name 'Shakespeare' isn't helping us here. Second, it's the most tickets we've ever sold to a show like this. Third, stop calling me Shirley."

At companies and organizations across the country, analytic dashboards give executives and Boards of Directors a performance snapshot. Dashboards often list sales results, projections and other key indicators, which are generally compared to this year's budget as well as totals from the same time last year. Additionally, and with a dashing stroke of creativity, the results might even be color-coded – green for 'everything's on track,' yellow for 'cautiously optimistic,' and red for 'problem area.' While this shorthand is helpful for time-starved observers who don't have their finger on the pulse of the organization, it presumes that the underlying models producing the colors are good or at least adequate.

And sometimes, it's not so easy to know how you are doing.

At the Cincinnati Shakespeare Company (CSC aka Cincy Shakes), a professional theater focused on Shakespeare and the classics, over 25,000 tickets are sold every year to 10 different productions, each with a typical run of 16-20 performances. As a small non-profit, these ticket sales are the lifeblood of the organization and a lot of effort is expended trying to understand how a given show is selling at

any point in time. At each staff meeting, the box office manager shares the latest sales data. At each Board meeting, the trustees review the dashboard with a specific interest in ticket sales.

But how can one tell if a show, which doesn't open for a month, will make its goal given its sales to date? Simply put, should it be red, yellow or green? More basically, how do you know if the original goal was even realistic in the first place? Despite having data records for every ticket sale of every performance of every show down to the minute of purchase for the past few years, the organization had no underlying reliable prediction model. The institutional gut had a single metric – "if we hit 20% of goal before our first performance, we will make the goal." Not only was this wrong about half the time, if a show failed to meet the rule-of-thumb, there was no time to do anything about it.

The CSC's Executive Director (and co-author of this article) realized that the theater possessed data that could be used more effectively to give advance notice of possible shortfalls or windfalls. He first enlisted one of his actors, albeit one with some quantitative background, to help improve the predictions. But the *ad hoc* methods they developed had deficiencies of their own, and so they reached out to Miami University, finally landing in the Department of Statistics. Over the next year and a half, a more systematic and principled approach was developed and deployed to the CSC, with accompanying interactive software. In this article, we describe this approach and application.

The value of good predictions is not simply the knowledge itself. The predictions promote better understanding which in turn enables better decision-making in the face of uncertainty and limited resources. Should a show be added, or a run extended, to provide additional supply for excellent demand? Should marketing support be increased to prop up sluggish sales, or should the theater simply play out the string and turn its attention to the next production in the queue? Is it going to be so bad that payroll is threatened?

While the choice of red, yellow, or green may not have mattered so much 400 years ago when Shakespeare's plays were first produced, they are crucial in today's world. Given the relatively recent

development in statistical modeling and information systems, surely (Shirley?) these tools could be leveraged to improve predictions and positively impact the company's success.

**2. Priming the Data for Action**

The raw data collected by the CSC's ticketing software includes the sorts of things one would expect: basic information about each show, patron information, date and time of all purchases, the number of tickets, and the amount paid for them (many of the tickets show an amount of 0 because they are passes issued for season subscribers). The data includes complete sales information for 40 shows from the last four seasons (2010/2011 to 2013/2014 which are Seasons 17-20 for the CSC). To construct the models we discuss in subsequent sections, we use all the shows in the database and make the simplifying assumption that the total sales of these shows represent independent observations. Because no major changes in, for instance, theater capacity have been realized over the years in question, the assumption of stationarity across seasons seems reasonable. Figure 1 shows a snapshot of the raw data, with personal data redacted.

| Bought Wh | Box Office | Show Name | Purchase Id | Person Id | Ticket Date | Purchase Date | Ticket Cc | Ticket Amount |
|---|---|---|---|---|---|---|---|---|
| tickets | 0 | A Man for All Seasons-18 | 16642111 | 2920746 | Thu 09/08/2011 07:30 PM | 8/9/2011 11:14 | 3 | 42 |
| tickets | 1 | A Man for All Seasons-18 | 16643710 | 2132064 | Sat 09/10/2011 07:30 PM | 8/9/2011 12:04 | 2 | 32 |
| tickets | 1 | A Man for All Seasons-18 | 16643984 | 1950720 | Fri 09/16/2011 07:30 PM | 8/9/2011 12:10 | 2 | 0 |
| tickets | 1 | A Man for All Seasons-18 | 16644171 | 1941384 | Sat 10/01/2011 07:30 PM | 8/9/2011 12:15 | 4 | 0 |
| tickets | 1 | A Man for All Seasons-18 | 16647091 | 1941619 | Fri 09/09/2011 07:30 PM | 8/9/2011 14:08 | 2 | 0 |
| tickets | 0 | A Man for All Seasons-18 | 16648085 | 2331059 | Wed 09/07/2011 07:30 PM | 8/9/2011 14:40 | 2 | 28 |
| tickets | 1 | A Man for All Seasons-18 | 16648442 | 2026746 | Sun 10/02/2011 02:00 PM | 8/9/2011 14:47 | 2 | 0 |
| tickets | 1 | A Man for All Seasons-18 | 16650801 | 1941374 | Sun 10/02/2011 02:00 PM | 8/9/2011 16:03 | 2 | 0 |
| tickets | 1 | A Man for All Seasons-18 | 16650879 | 2762905 | Thu 09/15/2011 07:30 PM | 8/9/2011 16:08 | 2 | 0 |
| tickets | 0 | A Man for All Seasons-18 | 16659862 | 1942254 | Fri 09/09/2011 07:30 PM | 8/10/2011 10:29 | 2 | 0 |
| tickets | 0 | A Man for All Seasons-18 | 16663879 | 2921919 | Sat 09/17/2011 07:30 PM | 8/10/2011 13:34 | 6 | 164 |
| tickets | 0 | A Man for All Seasons-18 | 16677809 | 1950718 | Fri 09/23/2011 07:30 PM | 8/11/2011 11:27 | 2 | 0 |
| tickets | 1 | A Man for All Seasons-18 | 16700009 | 2923803 | Thu 09/22/2011 07:30 PM | 8/12/2011 13:47 | 1 | 28 |
| tickets | 0 | A Man for All Seasons-18 | 16728421 | 2925097 | Fri 09/09/2011 07:30 PM | 8/14/2011 9:27 | 4 | 116 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15427361 | 2679244 | Tue 05/24/2011 07:30 PM | 5/24/2011 19:23 | 5 | 60 |
| tickets | 0 | A Midsummer Night's Dream(tour)-1 | 15429501 | 2679533 | Wed 05/25/2011 07:30 PM | 5/24/2011 22:23 | 2 | 24 |
| tickets | 0 | A Midsummer Night's Dream(tour)-1 | 15430640 | 2679627 | Wed 05/25/2011 07:30 PM | 5/25/2011 3:04 | 1 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15436977 | 1945047 | Wed 05/25/2011 07:30 PM | 5/25/2011 15:36 | 1 | 0 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15437178 | 2679974 | Wed 05/25/2011 07:30 PM | 5/25/2011 15:44 | 1 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15441262 | 1987410 | Wed 05/25/2011 07:30 PM | 5/25/2011 19:14 | 1 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15441286 | 2116964 | Wed 05/25/2011 07:30 PM | 5/25/2011 19:15 | 1 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15436683 | 2097310 | Wed 05/25/2011 07:30 PM | 5/25/2011 19:18 | 2 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15441540 | 1944141 | Wed 05/25/2011 07:30 PM | 5/25/2011 19:28 | 2 | 24 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15441556 | 2090108 | Wed 05/25/2011 07:30 PM | 5/25/2011 19:30 | 1 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 15441797 | 2090108 | Wed 05/25/2011 07:30 PM | 5/25/2011 19:47 | 1 | 12 |
| tickets | 1 | A Midsummer Night's Dream(tour)-1 | 21663412 | 2404278 | Tue 05/29/2012 07:30 PM | 4/18/2012 13:18 | 6 | 90 |

**Figure 1. A snapshot of raw data from CSC's ticketing software, with personal identifiers omitted.**

The sort of information we aimed to exploit for predictive purposes falls into two broad categories, shown in Table 1: (1) characteristic predictors and (2) sales-related predictors. Characteristic predictors describe inherent aspects of a show such as genre, time of year ("Interval" in Table 1; note that if a show straddles multiple categories it is assigned to that category in which it runs the most days), number of performances and whether the show has been staged over the time for which we have records (i.e. since "Season 17" or more recently). Genre is a manual classification into four general classes, including "Shakespeare Known" (famous plays such as *Midsummer Night's Dream*); "Shakespeare Obscure" (lesser known Shakespearean works such as *Richard II*); "Contemporary Classic" (a play a high school senior would know is not a Shakespeare and whose plot would be somewhat familiar); and "Contemporary" (typically by a 20$^{th}$ century playwright with a plot unfamiliar to the typical audience member).

**Table 1. Variables used to predict total sales**

| **Characteristic Predictors** | **Short Name** | **Description** |
|---|---|---|
| Number of Performances | np | Number of performances staged for a given show in a given season |
| Genre | genre | Contemporary (C)<br>Contemporary Classic (CC)<br>Shakespeare Known (SK)<br>Shakespeare Obscure (SO) |
| Interval | int | Spring (March 1 – May 31)<br>Summer (June 1 – August 31)<br>Fall (September 1 – November 20)<br>Holiday (November 21 – December 31)<br>Winter (January 1 – February 28) |
| Last Three Seasons | lts | 0 (including the present show, the show has not been staged more than once since the beginning of Season 17)<br>1 (including the present show, the show has been staged more than once since the beginning of Season 17) |
| **Sales-related Predictors** | | |
| Cumulative Sales | cs | Cumulative sales at $X$ days before/since first open (e.g. $X = -10$ means 10 days before first open; $X = 3$ means 3 days since first open) |
| Momentum | m | Sales between days $X - Z$ and $X$ |

The sales-related predictors in Table 1, Cumulative Sales and Momentum, are perhaps the most important. Cumulative Sales is defined as all sales of a given production from the beginning of the sale period through a chosen date $X$ days from first open (i.e. the first staging of the show). Momentum is calculated as the total sales in the $Z$ days preceding $X$. In general, we anticipate that the more cumulative sales at day $X$, and the greater the momentum, the higher the final sales will be.

To process the data from its raw form (Figure 1) to a structure that can be used to develop a predictive model, we adopt a simple approach: Take each complete show in the dataset as an independent observation, calculating the total sales for each show. To construct the final two sales-related predictors the analyst must specify two quantities, $X$ and $Z$, based upon the sort of predictions s/he would like to make. For instance, if a prediction of total sales is desired based upon cumulative ticket sales 10 days before first open and based on the sales (momentum) over the last 3 days, then $X = -10$ and $Z = 3$. This is illustrated graphically in Figure 2. The CSC is primarily interested in making predictions in the last month or so before the show opens, so $X$ is typically negative, though it could be specified as a positive whole number, representing the cumulative sales $X$ days into a particular show's run. The momentum, $Z$, on the other hand, only makes sense as a positive number.

Once $X$ and $Z$ are specified the cumulative sales and momentum predictors can be constructed from the data, and they are used along with the characteristic predictors given in Table 1 to develop a linear regression model (more on the choice of $X$ and $Z$, as well as model selection, in the next section). This model can then be used to predict the total sales for a future show. This approach also allows the Cincinnati Shakespeare Company to set an initial goal for sales while deciding which shows to stage for an upcoming season: they can simply use a model that uses only the characteristic predictors.

The general sales profile demonstrated in Figure 2 is typical. There are few sales until several weeks before the show begins, and shortly before the show opens, sales increase dramatically.

### 3. Constructing a Predictive Model that makes Good Predictions

Since the CSC runs 10 full shows every season, the models should automatically incorporate new data as it becomes available. Furthermore, since the theater may wish to make predictions at many different times (i.e. many different values of $X$) in the lead-up to various shows, the approach should be sufficiently flexible to allow this. For instance, the best model to predict total sales on the day the show opens (i.e. $X = 0$) using data available in 2015 might be quite different than the best model when a prediction is to be made 30 days ($X = -30$) in advance of first open using data available in 2018.

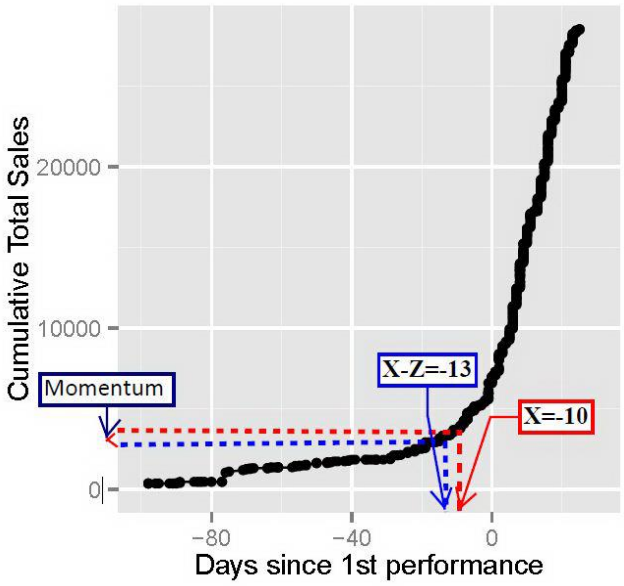


**Figure 2. Graphic representation of Cumulative Sales and Momentum, for a profile of ticket sales for A Man for All Seasons, Season 18 at the Cincinnati Shakespeare Company.**

Consequently, we use 10-fold cross-validation, in real-time, to select the predictive models. This means that the CSC can use an up-to-date raw dataset, and specify $X$ (and $Z$ in some cases), and the software will select an appropriate model and make the desired prediction (more on the software application later). Cross-validation is an approach that allows assessment of a model's quality in terms of its ability to predict new data. It proceeds by randomly partitioning the dataset into *k* mutually exclusive sets (folds), holding out a single fold, fitting a model to the remaining *k-1* folds, and using the fitted

model to predict the held out data. This process is repeated for each of the *k* folds. This procedure can also be used to choose a model (following Section 6.5.3 in *An Introduction to Statistical Learning with Applications in R* by James, Witten, Hastie, and Tibshirani). We considered for possible inclusion in our models all predictors in Table 1, as well as two-predictor interactions not involving Genre or Interval. Those interactions not considered were excluded to keep the models relatively simple, given that Genre and Interval have 4 and 5 categories, respectively.

As intimated above, there are two kinds of models of interest. First, for annual planning purposes the CSC would like a model without the sales predictors (see Table 1). For example, each year during budgeting time, the artistic director chooses a ten-show season. A key driver of the budget is the ticket projections for each show. Historically, the company would make educated guesses based on general sales trends ("Hey, this year we sold 20% more tickets than last year!"), institutional knowledge ("Audiences loved our last Hamlet, we did $35,000.") and learned heuristics ("History plays in May just don't sell in this city.") Using a model based on the characteristic predictors, the CSC can add these predictions to the matrix of inputs that they use to make decisions. Clearly this model will be limited in its capacity to predict, since the characteristic predictors represent only crude information. For the data at hand, the cross-validation procedure selected a model with just three terms, as can be seen in the "No sales info" row of Table 2. Thus, to predict the total sales of a show, you can use the following equation:

$$\begin{aligned}\text{Predicted Total Sales} = {} & 23,313 + 2,129(\text{Number of Performances} - 16) + 9,666(\text{Contemporary Classic}) \\ & - 5,703(\text{Number of Performances} - 16)(\text{Last Three Years})\end{aligned}$$

where "Contemporary Classic" is 1 if the show has that genre and 0 otherwise, and "Last Three Years" is 1 if the show has been staged more than once (including the current staging) since the beginning of Season 17. In Table 3 we give predictions for several shows using this and a more complicated model which we describe below.

**Table 2. Each row summarizes a model chosen via cross-validation. "CV RMSE" stands for Cross-Validation Root Mean Squared Error and is a rough estimate of the average prediction error for the model. "np" represents the Number of Performances predictor scaled so that 16 performances corresponds to $\mathbf{np} = \mathbf{0}$, "cs" the Cumulative Sales, "m" the Momentum, and "lts" indicates a play has appeared more than once since the beginning of Season 17 (see Table 1); "Spring" and "Summer" are indicator variables that specify whether a show is staged in the indicated interval, and "CC" indicates whether the show is a contemporary classic. Columns 3-14 are estimated regression coefficients of effects in the indicated model; a 0 indicates that the indicated predictor is not included in the model. For all models using sales data, Z=3.**

| Model Type | CV RMSE | Intercept | np | cs | m | Spring | Summer | CC | np*lts | lts*cs | lts*m | cs*m |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No sales info | 10,701 | 23,313 | 2,129 | 0 | 0 | 0 | 0 | 9,666 | -5,703 | 0 | 0 | 0 |
| $X = -30$ | 5,227 | 12,591 | 1,723 | 4.76 | 21.77 | -6,281 | 6,259 | 0 | 0 | 0 | 26.23 | -0.00435 |
| $X = -20$ | 6,116 | 18,002 | 1,497 | 2.27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| $X = -10$ | 5,361 | 12,018 | 1,568 | 3.04 | 0 | -4,763 | 5,673 | 0 | 0 | 0.64 | 0 | -0.00061 |
| $X = 0$ | 3,768 | 12,085 | 1,621 | 0.93 | 2.69 | -2,210 | 3,418 | 0 | 0 | 0 | 0 | 0 |

**Table 3. Total sales in dollars, and predictions, for six shows from Season 20 at the Cincinnati Shakespeare Company. For each show, the number of performances (np), the cumulative sales in dollars at $X = 0$ (cs), the momentum in dollars at $X = 0$, the show's genre, and in what part of the year it was staged (Interval) is given. Predictions are given using the $X = 0$ model as well as the model that uses no sales information. Note: Table 2 and the information in this table can be used to reproduce the predictions in this table, but due to rounding error they will not exactly agree.**

| Show | np | cs | m | lts | Genre | Interval | Total Sales | Prediction ($X = 0$) | Prediction (no sales info) |
|---|---|---|---|---|---|---|---|---|---|
| Hamlet | 20 | 15,464 | 4,560 | 0 | SK | Winter | 45,842 | 45,173 | 31,830 |
| The 39 Steps | 16 | 24,183 | 3,431 | 0 | C | Summer | 54,166 | 47,139 | 23,313 |
| Rosencrantz and Guildenstern are Dead | 16 | 18,257 | 1,938 | 0 | CC | Winter | 32,910 | 34,211 | 32,978 |
| Of Mice and Men | 15 | 7,867 | 1,755 | 0 | CC | Fall | 24,025 | 22,478 | 30,849 |
| Twelfth Night | 14 | 12,244 | 2,907 | 0 | SK | Holiday | 28,200 | 28,014 | 19,054 |
| Oliver Twist | 16 | 8,011 | 2,744 | 0 | CC | Fall | 24,270 | 26,897 | 32,978 |

The second model of interest, for marketing purposes as well as for cash flow management, is one which includes sales predictors to help understand how a given show is selling at any point in time throughout the sale period. The earlier the company can spot a potential shortfall, the more they can impact it with additional marketing spending. Managerially, knowing how a given show is doing at any point in time can help reduce leadership and board concern, or at least manage their expectations and avoid unpleasant surprises.

Figure 3 gives visual results of the cross-validation for several different values of $X$, chosen as representative of $X$'s that interest the CSC. (The number of days over which Momentum is calculated has been specified as $Z = 3$ throughout, based on some initial testing that suggested it performed better than $Z = 7$. These two values for $Z$ were chosen because of their pleasing intuition; more extensive testing could be performed.) Note that the rough estimate of the average prediction error, "CV RMSE", is 2-3 times larger for the model that does not use sales data. It is clear that not only does incorporating sales information make predictions better, but also that using sales information as late as the opening day of the show improves the predictions even further. On the other hand, there doesn't appear to be a great difference between 10, 20 and 30 days.

The cross-validation procedure explores models with a various number of terms, and looks for the model size that best balances underfitting and overfitting, as evidenced by the smallest average root mean squared error. Cross-validation plots, like those in Figure 3, typically have a U-shape, indicating a sweet spot in terms of the number of predictors. Our plots do not uniformly conform with that ideal. This is somewhat unusual, but may be partially explained by the relatively small number of observations and the categorical nature of some of the predictors. Indeed, when the procedure is run with only the numeric predictors Cumulative Sales, Momentum, and Number of Performances and their associated two-predictor interactions, the plots are better—though still not perfectly—behaved.

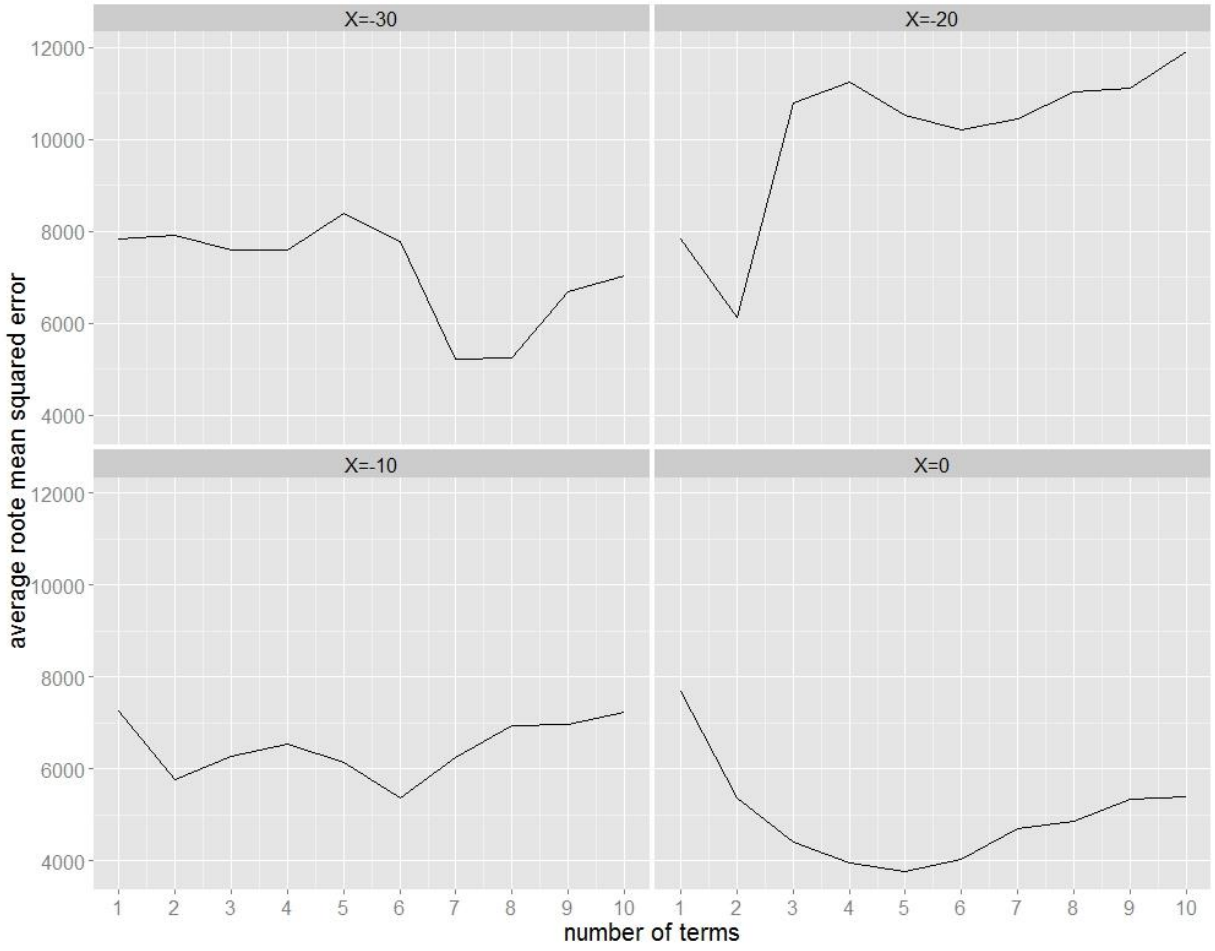


**Figure 3. Results of the model selection by 10-fold cross-validation. The y-axis gives a measure of the average error; the x-axis gives the number of terms in the associated model.**

Consider the predictions made by the $X = 0$ model for shows identified in the dataset as Season 20 in Table 3. Overall, they appear quite good and certainly much better than the predictions from the model that does not use sales information. This is not unexpected, given that the average root mean squared error of this model is so much smaller.

## 4. What Are the Most Important Predictors?

From Table 2, one can see that the number of performances ("np") has a reasonably stable relationship with the total sales for the models that incorporate sales information. That is, for the

models that use sales information, each additional performance is associated with a roughly $1,500-1700 increase in total sales. Furthermore, it is clear that Spring shows don't tend to do as well, and Summer shows experience a bump. Anecdotally, the 'spring dip' may be due to the many artistic and entertainment alternatives in the region that emerge when the weather breaks, from professional baseball games to end-of-year high school musicals, graduation parties, etc. Similarly, the 'summer fling' may be driven by the benefit of being the only theatrical game in town during that part of the year.

For the range of X's considered, the magnitude of the association between cumulative sales and total sales typically and expectedly declines as we use more recent information. For instance, an increase of one dollar of cumulative sales 20 days out is associated with an increase in total sales of around 2.3 dollars (holding other predictors constant), whereas that factor has shrunk to 90 cents for the $X = 0$ model. Note that for the $X = -30$ model, the cumulative sales coefficient (4.8) isn't quite as large as it appears because also included in this model is an interaction between Cumulative Sales and Momentum. For a show that has a momentum of $279 (roughly the average momentum when $Z = 3$), the cumulative sales factor is reduced to $4.8 - 0.0044(279) = 3.6$, which is in line with the other models.

Note also that momentum plays a role in three of the four models that use sales information. For the $X = -30$ and $X = -10$ models, the effect of momentum is obscured by its involvement in interactions, but by examining the $X = 0$ model we can interpret its effect in a straightforward way: That is, for every additional dollar sold in the last three days, a $2.7 increase in total sales is expected assuming all other predictors are held constant.

**5. Toward a Dashboard: A *Shiny* Application**

As described in the Introduction, a dashboard is crucial for an organization like the CSC because the primary stakeholders are business and theater professionals, not statisticians. Consequently,

unfiltered “regression output” or “R code” will obfuscate rather than clarify. A good underlying predictive model is necessary for the happy melding of analytics and business, but it is not sufficient. The results should be presented in a way that engages and highlights all the results that are important to the organization.

As described in Section 2, a significant amount of data reformatting is required to get from the raw database to a form amenable to modeling. The CSC’s limited resources prevent them from hiring someone to build a modelling platform in-house. This necessitated a stand-alone, user-friendly software solution. Accordingly, we used RStudio’s Shiny R package, which facilitates the construction of interactive statistics applications. Behind the scenes, the raw data are processed and combined with user input to produce plots, models, and predictions. The interface displayed in Figure 4 is designed to meet a wide range of needs for those monitoring ticket sales at the CSC. The functions described below are based upon the modeling approach described in Sections 2 and 3. That is, linear regression models are chosen via cross-validation and deployed to make predictions in several different ways.

Firstly, any show represented in the database, whether complete or in-progress, can be plotted (e.g. Figure 4). This allows the analyst to quickly visualize the status of a particular show.

Secondly, the analyst can fit a model and see the results for all completed shows. The models can either exclude or include sales information, and in the latter case the type of model that is fit can be controlled by specifying X and Z. This gives an overall view of a model and its predictions for past shows. It can also prompt questions that increase understanding. For instance, a Season 17 production, “Frankenstein: The Modern Prometheus”, is consistently overpredicted by the model. In this particular case, Frankenstein was a one-man, off-night show paired with Dracula, and likely had less publicity than a typical show. Thus, the lower-than-expected sales are not too surprising.

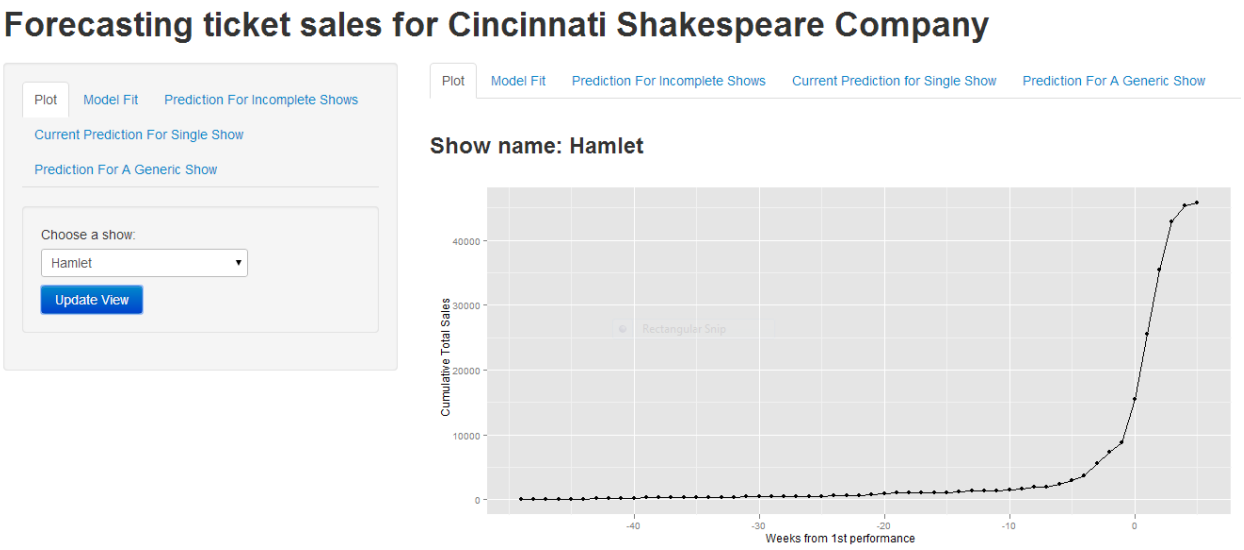


**Figure 4: Shiny application user interface.**

Thirdly, the program can make predictions for all incomplete shows in the database, based on the last sale date in the database for each show. For instance, in the database used in the development of this application, tickets for "Henry IV" were being sold. Based on the date of the last ticket sale on record there were still 9 days until it opened. Thus, this functionality constructs a model to predict the show's total sales based on its current sales, its momentum (Z=3), and the characteristic predictors given in Table 1. This is accomplished, of course, by using $X = -9$ and constructing a model using all of the completed shows in the database. This part of the program produces sales predictions for all incomplete shows in the current database within 60 days of first open. On the downside, it requires an updated database, which, since it is fairly large, cannot be obtained immediately.

Since the CSC would like to have feedback every day, the fourth function of the program allows the user to choose any show in the database within 60 days of opening and specify its current sales and momentum (Figure 5). Then, based upon the current date and the date of the show's first performance, a prediction will be made. This is perhaps the most important functionality because it does not require

an up-to-date database but still allows, as long as the user can specify the current sales and recent sales, an up-to-date prediction.

Finally, the fifth function of the program is to allow a prediction to be made for a generic show. This show does not have to be found in the database, so this function can be used for planning purposes. It requires the specification of the show's name, its genre, the time-of-year it will be staged, the number of anticipated performances, and whether the show has been staged before. This information is enough to allow a prediction, but of course current (or expected) sales data can be used as well. The primary benefit of this function is to allow goals to be made for shows with various characteristics that are being considered for upcoming seasons.

Note that for each of these modeling functions, the analyst can specify whether or not to include current sales information (Current Sales and Momentum) in the model. For instance, in the "Prediction for Incomplete Shows" function, if current sales information is not used the program will make a prediction of all incomplete shows in the database based only on characteristic predictors.

We note that the Shiny application is fully functional for the CSC, but has several deficiencies that distinguish it from fully tested, commercial software. For instance, occasionally an error surfaces that prevents the cross-validation procedure from successfully constructing a model. In this case, the user should restart the application to obtain the desired predictions. There are also some quirks in the GUI: for instance, sometimes the output columns wrap-around in an unattractive way, and the input and output panels don't automatically coordinate with each other. Though the CSC indicates that these problems with the GUI are not problematic for their use of the application, newer versions of the application might address these issues.

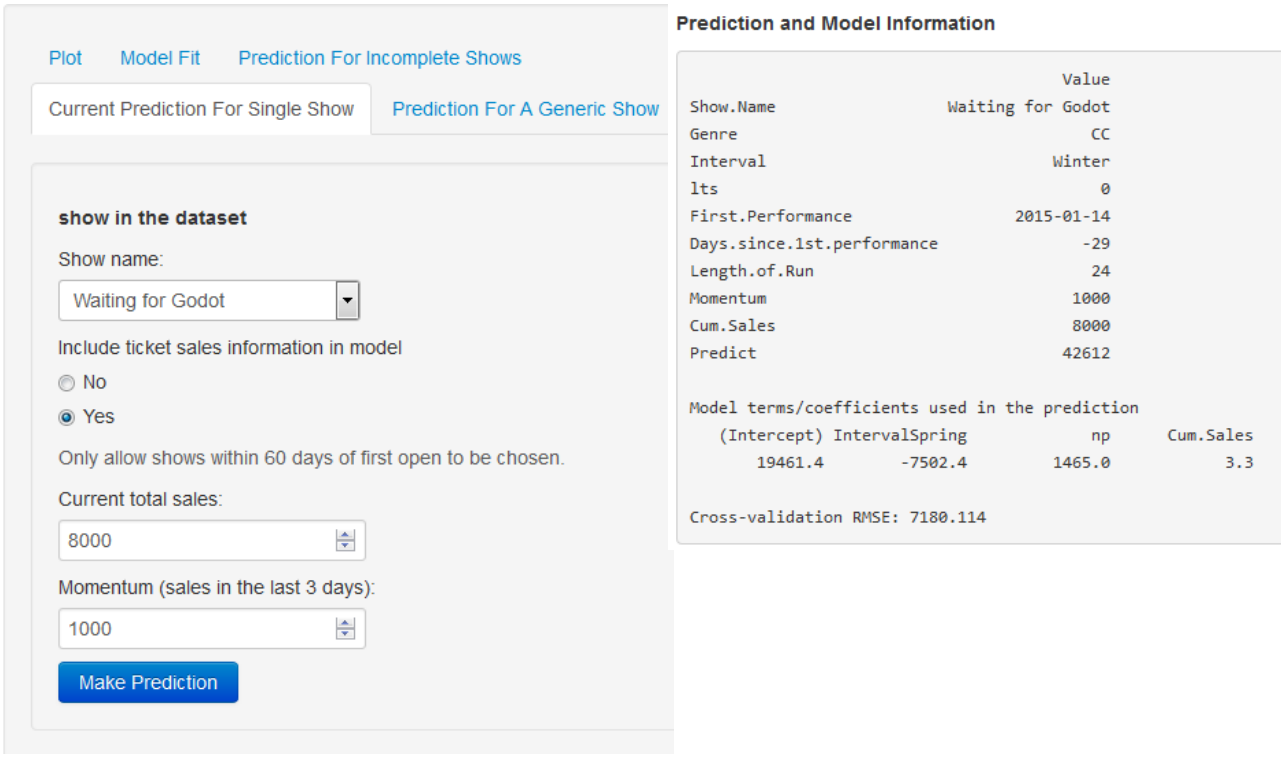


**Figure 5. Current Prediction For Single Show (left, input; right, partial output).**

## 6. Summary and Discussion

Information about key performance indicators is important in the operation of any organization, and when information can be exploited to make accurate predictions about the future, it is invaluable. The more effective the models underlying the information and predictions, the more informed, flexible, and empowered the decision-makers are. In this particular case, we have developed least-squares regression models to make predictions to serve a small, non-profit theater in Cincinnati. The models have been deployed to the Cincinnati Shakespeare Company in the form of a Shiny application, and are being used by the organization.

Consider an example of the CSC's use of these predictive models. In March of 2014, there were two shows remaining in the CSC's season, Two Noble Kinsmen (TNK) and Private Lives (PL). The question confronting the theater was where to invest the remaining marketing dollars. At the time, the heuristics

used in the past, based on a dataless rule of thumb, suggested that TNK was lagging and PL was on track. The model, however, projected that TNK would meet goal while PL was trailing fairly significantly. This caused a shift in the CSC's thinking, and instead of spending on TNK, they shifted their money to PL. In the end, the PL's momentum was increased and both shows ended up exceeding their goal. This anecdote illustrates how data, as analyzed via a predictive model, has provided insight and altered the behavior of decision-makers within this organization.

Ultimately, the CSC is interested in two sorts of questions. First, questions about turning data into predictions: How can one set realistic goals for a show in the planning stages? How can one tell at a given point in time whether sales are tracking with the goal or coming up short? Second, there are questions of a more prescriptive nature: Can predictions be used to drive organizational decision making, for instance extending a run if demand is strong or supporting additional marketing to boost lethargic sales? The model developed can directly answer the first kind of question, and allows the CSC to further explore the second.

It is likely that further improvements could be made to these methods. The current procedure doesn't explicitly use all of the data for each show. That is, once the cumulative sales and momentum, indicated by X and Z, are specified, the construction of the model ignores the rest of each show's sales profile. It might be better to model these profiles directly. This would, of course, require more complicated statistical methods. One could also imagine other predictors that, if they were available, would probably improve the predictions. For instance, quantifying the advertising for a particular show at a given X, or including an external entertainment index to quantify the entertainment options in Cincinnati at a particular time.

With the model and software described in this article, however, the Cincinnati Shakespeare Company is now better equipped to make principled predictions and use them to take appropriate actions. Several times a week, the leadership team hears how the sales are doing for each upcoming

show. The dashboard the CSC uses now has an effectively color-coded green each time the model predicts the show will make its goal, yellow when the model shows it slightly under and red when things look bleak. These cues, even when they are 10, 20, or even 30 days out are equipping the organization with insight. Now, the actor can know when his presence and performance is inspiring both awe and healthy ticket sales, the Board member doesn't freak out except for a very good reason, and the marketing professional doesn't spend money on advertising unless the data call for it. And now, they can all walk into the bar and have a celebratory drink as the thinking has made it so.

**Further Reading**

James, Witten, Hastie, and Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Shiny (2013). RStudio: Integrated development environment for R (Version 0.98.501) [Computer software]. Boston, MA. Available from http://www.rstudio.org/.