Communications of the IIMA

Volume 12 | Issue 1 Article 2

2012

The JMP to Catalina Island Competition

Harold Dyck
California State University, San Bernardino

Xinran Wang
California State University, San Bernardino

David Kung
University of La Verne

Frank Lin
California State University, San Bernardino

Follow this and additional works at: https://scholarworks.lib.csusb.edu/ciima

Recommended Citation
Dyck, Harold; Wang, Xinran; Kung, David; and Lin, Frank (2012) "The JMP to Catalina Island Competition," *Communications of the IIMA*: Vol. 12: Iss. 1, Article 2.
DOI: https://doi.org/10.58729/1941-6687.1178
Available at: https://scholarworks.lib.csusb.edu/ciima/vol12/iss1/2

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Communications of the IIMA by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.

The JMP to Catalina Island Competition

Harold Dyck
California State University, San Bernardino, USA
hdyck@csusb.edu

Xinran Wang
California State University, San Bernardino, USA
Wangx310@gmail.com

David Kung
University of La Verne, USA
kungd@ulv.edu

Frank Lin
California State University, San Bernardino, USA
flin@csusb.edu

ABSTRACT

This paper describes a forecasting competition developed for a graduate course in information decision management to illustrate a data mining technique. The competition was inspired by the Netflix Prize competition (Bennett & Lanning, 2007) and other crowdsourcing competitions. A dataset of 276 monthly observations from 1989 through 2011 was partitioned into a learning dataset of 264 observations and a holdout sample of the 12 observations for 2011. After being subjected to a teaching module covering time series forecasting methods, teams of students tried their luck forecasting the holdout sample with an objective of minimizing the root mean square error. We learned that combining models results in a 7% decrease in RMSE.

Keywords: Crowdsourcing, classroom competitions, forecasting

INTRODUCTION

The Netflix Prize (Bennett & Lanning, 2007) was a one million dollar prize offered by Netflix to the team able to reduce the root mean square error in its recommender system by 10%. The excitement generated by the prize is reflected by the announcement of the winner on the front page of the *New York Times* (Lohr, 2009). Over a three-year period, the team "BellKor's Pragmatic Chaos" was able to beat over 50,000 competitors from around the world using data mining algorithms.

This paper describes a competition inspired by the Netflix Prize for a graduate information decision management course taught in winter 2012. A dataset of 276 monthly observations of Catalina Island cross-channel visitors was obtained from the Catalina Island Chamber of Commerce & Visitors Bureau. Because students were encouraged to use the statistical program

JMP (2012), we called the competition the "JMP to Catalina Island Competition." The competition was used to illustrate a data mining technique as one of the objectives for the course.

THE VALUE OF COMPETITIONS IN THE CLASSROOM

There exists a bit of literature on the use of competition to promote learning in the classroom. J. R. Anderson (2006), for example, looks at the use of competitive and cooperative approaches to motivate students. He concludes that a balance approach is best. Bandura (1977) created a theory of social learning: we learn through observing others. But the state of online competitions is just beginning and there does not exist a large body of knowledge. But the stage is set. A new website, Kaggle.com, has been set up to help instructors create online competitions and may provide a potentially rich opportunity for further study. The site describes several other crowdsourcing competitions. This site was introduced to one of this paper's authors at the 2012 Joint Statistical Meetings. For more on this subject, see Elkan (2012), Gonzalez-Brenes (2012) and Sonas (2012).

DATA MINING

Data mining "is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand, Mannilla and Smyth, 2001, p. 1). Data mining is a convergence of statistics, computer science and database principles. A good description of the data mining process can be found at SAS.com (n.d.) with their SEMMA methodology: sample, explore, modify, model, assess. That is, one first begins by sampling from the dataset and partitioning it into a training, validation and test set. Exploratory data analysis is then performed using data visualization techniques and other simple techniques. The data may then be transformed through normalization or linear transformations. Modeling the data is typically accomplished through computer intensive methods. Finally, the models are assessed by their ability to predict the data in the test set. This data mining process can be contrasted with the traditional hypothesis testing methods of statistics where the test is stated *a priori*.

JMP

JMP is an interactive, visual statistics package that incorporates many data mining techniques, (including neural nets, classification and regression trees), as well as more traditional techniques such as regression, ANOVA and time series methods. It also has modules for design of experiments (DOE), quality control (control charts, Pareto plots, cause and effect diagrams, Taguchi analysis and capability indexes) and survival analysis. JMP is a subsidiary of SAS and has the support and training SAS is known for, including live and on demand Webcasts, seminars, conferences, textbooks, and more. Students can download a free 30-day copy or pay $30 or $50 for six or twelve month access. A more limited version can be obtained for around $10 and bound with a text. An instructor considering using JMP in the classroom should start at JMP.com. A list of textbooks using JMP can be found at that site. One of the authors of this

paper has successfully used JMP in business statistics courses as well as the DSS course mentioned in here.

THE COMPETITION

To set up the competition, the first step was to obtain a dataset and partition it into the learning set given to the students. The classroom instruction began with some basic training in the use of JMP, but students were encouraged to make use of the beginner's tutorial included with the software. Examples using time series analysis were presented and the rules of the completion were spelled out. The game evolved as more feedback was obtained. For example, we started requiring the parameter estimates be given in a certain format so that the model was completely specified and forecasts could be reconstructed. In addition, two years of forecasts were required so that we could provide useful forecasts into the actual future (as opposed to just the hold-out sample).

The goal of this competition is to discover the best seasonal autoregressive integrated moving average (ARIMA) model based on the root mean square error (RMSE) using the holdout sample of 12 monthly observations of 2011. In so doing, it is hoped that this competition gave students an understanding of how information tools help businesses make better decisions. Further, the results of our predictions for 2012 will be shared with the Catalina Island Chamber of Commerce for their planning purposes.

DATA AND METHODOLOGY

The competition is based on Catalina Island cross-channel carrier monthly visitor count report of 276 monthly observations from 1989-2011 provided by the Catalina Island Chamber of Commerce & Visitors Bureau. Although the Chamber also provided counts of cruise ship visitors, we decided to concentrate on the cross-channel visitors. It was felt that aberrations in counts of cruise ship visitors due to the 2009 swine flu epidemic made time-series modeling problematic. The data was partitioned into the first 264 observations to be used as the learning data and the 12 observations of 2011 used as a holdout sample. This holdout sample was not given to the students. Students were instructed in Box-Jenkins ARIMA methodology and seven teams of two students each were given the task of finding a model to submit to the instructor over about a 4-week period. At the end of each week a leaderboard was published showing the results of the teams' forecasts based on RMSE.

The competition ran from February 20 through March 18. Participants were encouraged to use the statistical package JMP to predict the Catalina Island cross-channel carrier monthly visitors from 2011 through 2012 (the actual 2012 values were unknown at the time of the competition) and submit their two-year prediction of results and model information by midnight on Friday each week. The instructor then evaluated the prediction results based on the RMSE over 2011and announced the team ranking on the leaderboard at the end of each week. Figure 1 shows a JMP plot of the data, which obviously follow a highly seasonal pattern. Visitor counts range from 10,585 to 121,064. JMP is a highly interactive, visual and dynamic statistical package available

at JMP.com. Since JMP is a SAS product and has substantial support, we had the good fortune to have two representatives from JMP visit our campus and give a two-hour demonstration of their software. The package also has substantial online support. See, for example, *http://www.jmp.com/support/help/Seasonal_ARIMA.shtml*.

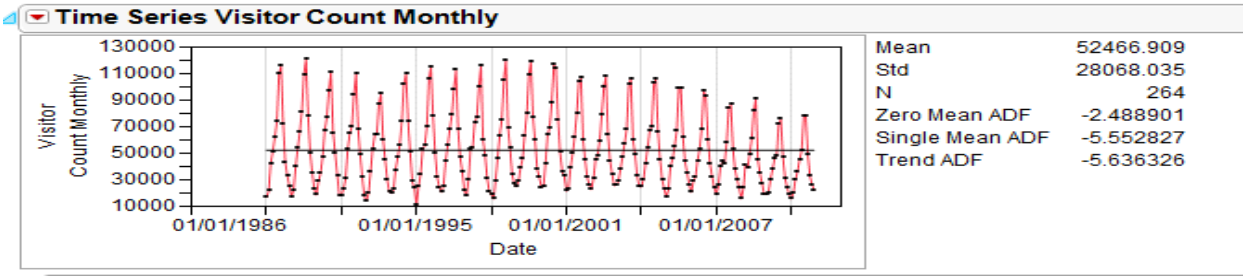

Figure 1: Seasonal ARIMA Modeling and RMSE Methods.

Although the teams mostly used seasonal ARIMA modeling to find the best model, they were allowed any method they wanted, including other time series approaches such as moving averages, exponential smoothing and decomposition methods. Students were taught the following process of model identification, estimation, adequacy checking and forecasting suggested by Box and Jenkins (1976). A description of how the students used the JMP software and identified their models follows.

Model Identification

By using time series analysis in JMP with data entered, students will first observe the autocorrelation function (ACF) and the partial autocorrelation function (PACF) of the visitor data. Figure 2 shows the spikes beyond the 2 standard error lines where something needs to happen. The goal is to try different models until the model is judged adequate.

The ACF and PACF suggest which models to try. We identify the model as an ARIMA (p,d,q) model, where p, d, and q are the integers representing the highest order of the autoregressive, integrated and moving average terms of the model. The seasonal ARIMA models are denoted as Seasonal ARIMA(p,d,q)(P,D,Q) ^{s}, where "p", "d" and "q" are refer to autoregressive, differencing, and moving average terms of ARIMA model's parameters, respectively, "s" is the number of periods per season. "P", "D" and "Q" are elements for specifying the seasonal autoregressive order, P, seasonal differencing order, D, and seasonal moving average order, Q.

Typically, the first step is to determine the order of differencing, d. If the plot of the data shows a linear trend, first order differencing may be appropriate, that is d=1. Another way to determine the order of differencing is by observing the ACF to see if it dies out slowly. ARIMA(0, 1, 0) just means first order differencing. JMP calls this model I(1). ARIMA(1, 0, 0) is referred to as AR(1). It is a first order autoregressive model. If y_t is our original series, then first order differencing can be represented by:

$(1-B)^1 y_t = y_t - y_{t-1}$,

Where B is the backshift operator, such that $By_t = y_{t-1}$. The d^{th} order differencing (and D^{th} order seasonal differencing) is $(1-B)^d(1-B^s)^D y_t$. The idea is to determine what d and D are to achieve stationarity. Usually, *d* is 0 or 1. JMP uses the notation $w_t = (1-B)^d y_t$, where w_t is our transformed variable after differencing.

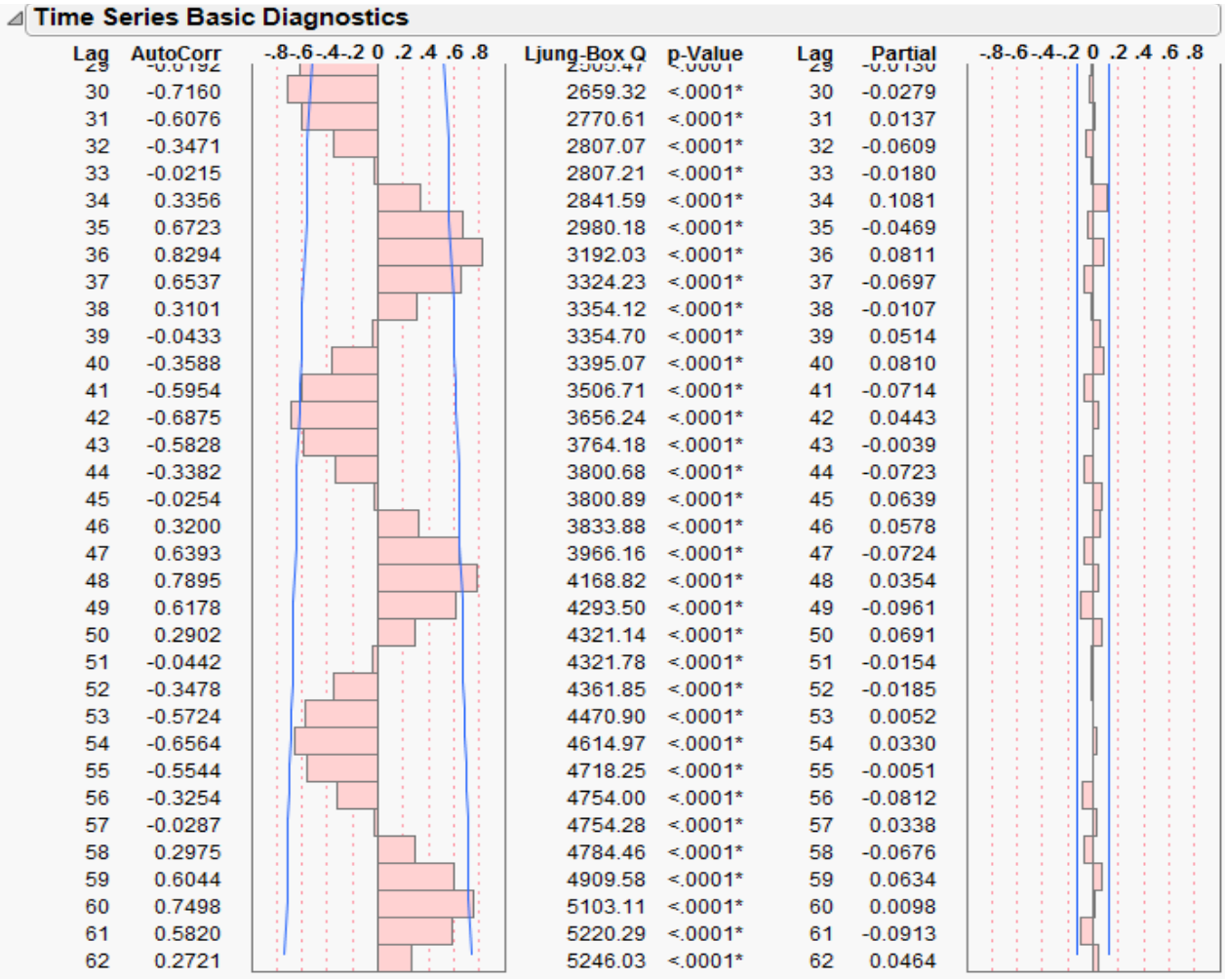

Figure 2: Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).

An autoregressive model is written as AR (p) or ARIMA (p, 0, 0). A first order autoregressive model, or AR (1), can be written as:

$$y_t = \delta + \phi_1 y_{t-1} + a_t$$

This is just our old regression model, except that we're regressing on the same dependent variable lagged by one period. The δ is our constant term, ϕ_1 is our slope parameter and a_t is our error term. Note that the above equation can be written as $(1- \phi_1 B)y_t = \delta + a_t$ or $\phi_1(B)\, y_t = \delta + a_t$, where $\phi_1(B)$ is a first-order polynomial in *B*. Thus, for AR (1), there is only one autoregressive term, ϕ_1, to estimate. For AR (p), there would be p autoregressive terms to estimate and, in general,

$\phi_p(B)\Phi_P(B) = (1- \phi_1 B- \phi_2 B^2- \phi_3 B^3 \ldots- \phi_p B^p)(1-\Phi_1 B-\Phi_2 B^2-\Phi_3 B^3- \ldots -\Phi_P B^p)$

Where $\phi_p(B)$ is the non-seasonal auto regressive polynomial of order p, and $\Theta_P(B)$ is the seasonal auto regressive polynomial of order P. Similarly, we have moving average polynomials of order q and Q:

$\theta_q(B)\Theta_Q(B) = (1- \theta_1 B- \theta_2 B^2- \theta_3 B^3 \ldots- \theta_q B^q)(1-\Theta_1 B-\Theta_2 B^2-\Theta_3 B^3- \ldots -\Theta_Q B^Q)$

After model identification, estimation and adequacy checking, students made their forecasts and submit them to the instructor. The instructor then calculates the RMSE and ranks the prediction of each group based on their seasonal ARIMA modeling results, degrees of freedom and actual data of Catalina Island cross-channel carrier monthly visitor on 2011, using:

$$\text{RMSE} = \sqrt{\frac{\sum(y_t - \widehat{y_t})^2}{dof}}$$

Where y_t is the actual monthly visitor data on 2011; $\widehat{y_t}$ is the predicted monthly visitor data on 2012; dof, or degrees of freedom, is the sample size (264 in our case) less the number of parameters estimated. The Figure 3 shows the model information and RMSE result for the best model of each of the seven groups.

Rank	RMSE	Group	Model
1	1208.7178	Group 3	(3, 1, 3)(24, 0, 24)^12
2	1235.154	Group 4	(2, 3, 3)(3, 2, 2)^12
3	1288.3789	Group 3	(3, 1, 3)(12, 0, 12)^12
4	1394.8217	Group 2	(3,0,3)(3,0,3)^12
5	1410.4884	Group 4	(1, 3, 3)(3, 2, 1)^12
6	1624.9991	Group 2	(3,0,3)(3,1,3)^12
7	1637.1773	Group 6	(3,1,3)(3,4,3)^12

Figure 3: Ranking of the Seven Student Groups

The Combined Model

After the competition was over the winning team was presented with the JMP to Catalina Island prize (baseball caps with the Catalina Island logo). The instructor then demonstrated the advantage of combining models using the top seven models submitted.

The **combined model by** following the steps given below was found by weighting the forecasts from seven models using a weighting scheme based on the RMSE of each model. Thus,

$$w_i = \frac{\frac{1}{RMSE_i^2}}{(\frac{1}{RMSE_1^2} + \frac{1}{RMSE_2^2} + \cdots + \frac{1}{RMSE_7^2})}$$

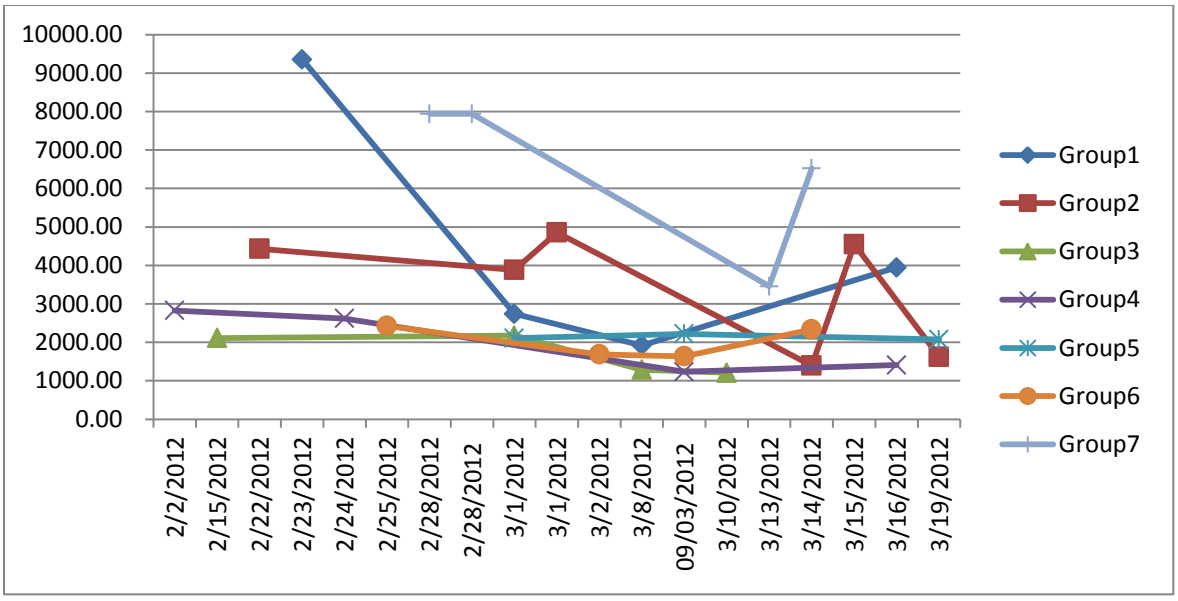

Figure 4: Timeline Showing How Each Group Performed Over Time.

We next calculate the combined forecast at time t using the above weights:

$\hat{y}_{ct} = \sum_{i=1}^{7} w_i \hat{y}_{it}$, and calculated the weighted degrees of freedom of the combined model using:

$$Wtd.dof = \frac{dof_i}{w_i}$$

Finally, we estimated the root mean square error (RMSE) for combined:

$$\text{RMSE}_c = \sqrt{\frac{\sum_{t=1}^{12}(y_t - \hat{y}_{ct})^2}{Wtd.dof}}$$

The result shows that root mean square error (RMSE) for combined model is 1,123.13, which is about 7 % smaller than the best participant's model RMSE result. The comparison of the combined forecast model and best participant's forecast model is shown in Figure 5. Figure 6 graphs the combined model and best student model with actual data.

Rank	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Group	Group 3	Group 4	Group 3	Group 2	Group 4	Group 2	Group 6
RMSE	1,208.7	1,235.2	1,288.4	1,394.8	1,410.5	1,625.0	1,637.2
MSE	1460999	1525605	1659920	1945528	1989477	2640622	2680350
1/MSE	6.84E-07	6.55E-07	6.02E-07	5.14E-07	5.03E-07	3.79E-07	3.73E-07
Weights	0.184	0.177	0.162	0.139	0.135	0.102	0.101
wtd dof	38.55	39.92	38.96	34.77	30.88	24.39	20.41

Figure 5: Calculating Weights and Dof for the Combined Model.

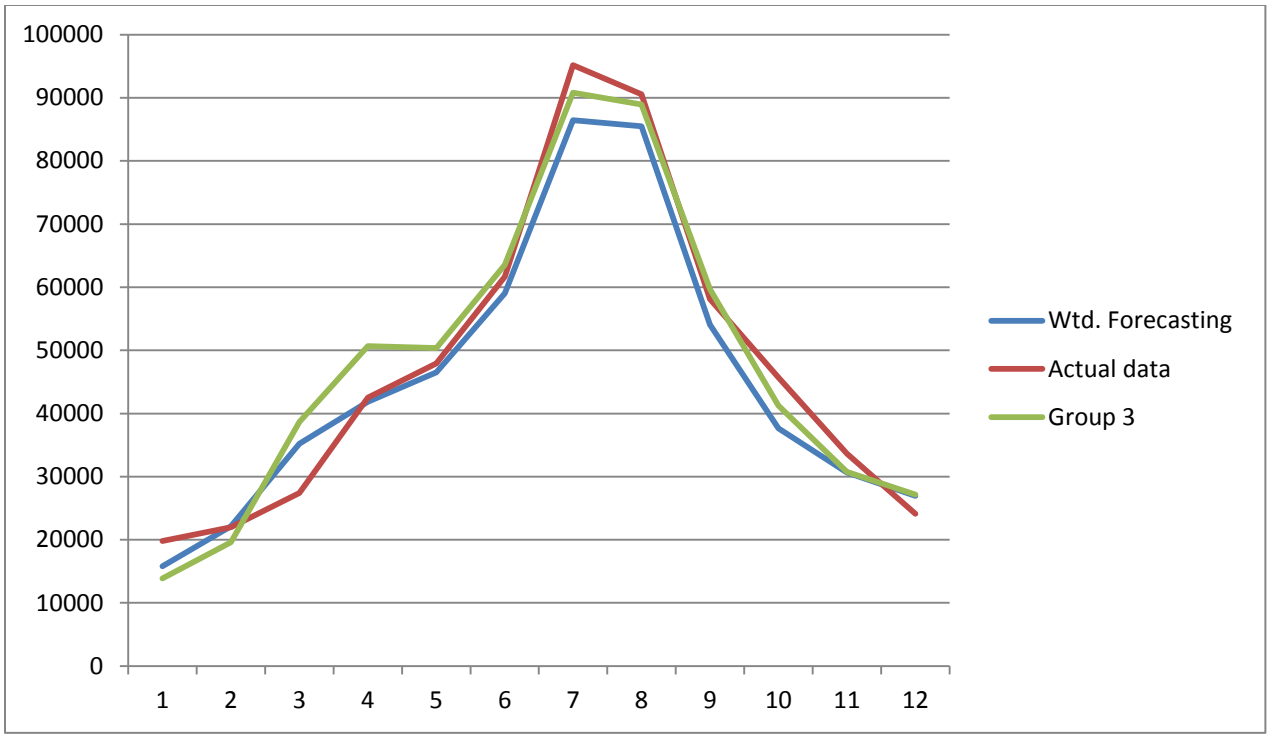

Figure 6: Comparison of Combined Model and Best Student Model with Actual Data

CONCLUSIONS AND DISCUSSION

This paper describes the use of a Netflix Prize-type competition in a graduate course in information decision systems. Students were exposed to the data mining concept, learned a specific technique and gained experience with the JMP statistics program from SAS. Students agreed that it was a useful exercise and appreciated the hands-on access to real data and a real problem.

We think that it might be interesting to modify the competition by including the addition of independent variables. For example, Google Insights for Search allows time series to be downloaded with normalized volumes of terms searched (such as the term "Catalina Island"). This was considered in the present case but we decided to keep things simple in this first iteration. Additionally, there exist other weighting schemes for the combined model, such as using the AIC or BIC.

REFERENCES

Anderson, J. R. (2006). On cooperative and competitive learning in the management classroom. *Mountain Plains Journal of Business and Economics, Pedagogy, 7*, 1-10.

Bandura, A. (1977). *Social learning theory*. New York, NY: General Learning Press.

Bennett, J., & Lanning, S. (2007). The Netflix prize. *Proceedings of KDD Cup and Workshop 2007*. Retrieved August 1, 2012, from: *http://www.netflixprize.com/assets/ NetflixPrizeKDD_to_appear.pdf*

Box, G. E. P., Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*, San Francisco, CA: Holden-Day Inc.

Elkan, C. (2012). *Data analysis contests: From good to better?* Paper presented at the 2012 Joint Statistical Meeting, San Diego, California.

Gonzalez-Brenes, J. P. (2012). Predicting travel times for the M4 highway in Sidney: How we won Kaggle.com's first $10,000 data mining challenge. Paper presented at the 2012 Joint Statistical Meeting, San Diego, California.

Hand, D., Mannila, H., & Smyth, P. (2001), *Principles of data mining*. Cambridge, MA: Massachusetts Institute of Technology.

JMP. (2012). *Modeling and multivariate methods, Performing time series analysis, Seasonal ARIMA*. Retrieved August 1, 2012 from: *http://www.jmp.com/support/help/Seasonal_ ARIMA.shtml*

Lohr, S. (2009, June 26). And the winner of the $1 million Netflix prize (probably) is…. *NY Times*, p.1A. Retrieved from http://bits.blogs.nytimes.com/2009/06/26/and-the-winner-of-the-1-million-netflix-prize-probably-is/

SAS. (n.d.). Predictive analytics and data mining. Retrieved from http://www.sas.com/ technologies/analytics/datamining/index.html

Sonas, J. (2012). Using a crowdsourcing contest to find a more predictive chess rating system: The Deloitte/Fide chess rating challenge. Paper presented at the 2012 Joint Statistical Meeting, San Diego, California.

This page left intentionally blank