Paper SAS4612-2020

Turning the Crank: A Simulation of Optimizing Model Retraining

David R. Duling, SAS Institute Inc.

ABSTRACT

Model retraining is a common practice in the advanced model life cycle. However, the critical question is how do you know when you need to retrain the model? Once the model is retrained, how do we determine when we need to redeploy the model? Can we predict how long the model will be relevant? The answers can depend on one or more of many factors including calendar fluctuations, business cycles, data drift, model performance, expected benefit, and many others. Given those factors, we want to find the optimal points in time to retrain and redeploy a predictive model. This paper presents a simulation study of different strategies and techniques for optimizing model retraining with the goal of maintaining optimal business performance.

INTRODUCTION

Most data mining studies focus on building the most accurate predictive models. Competition programs such as Kaggle often supply a single large data set and pose a unique prediction problem. The typical task is formed to create one predictive model with maximum test data accuracy. Competitive models are often formulas that have been carefully tuned to the unique objective function on the single large data set. Once the competition is completed, the supplier of the data harvests the knowledge created by the competitors. The competitors move on to the next challenge. However, data does not exist as a single point in time. In real-world applications, data is continuously collected from operational systems and is subject to changing conditions. The data collected in the second month may be different than the data collected in the first month. Therefore, we may need to create a new model in the second month or later. The process of creating a new model to adapt to changing patterns in the data is called "model retraining". This paper expands on a sample of retraining strategies using a long running data sample from a publicly available source.

MODEL DECAY

In our 2019 paper "The Aftermath What Happens After You Deploy Your Models and Decisions", we described how models are scored in an operational process. We also concluded with a section on model decay and retraining, and then presented a theoretical example. Figure 1 shows two plots from that paper. In both plots, the lower green line shows a measure of model performance for a real model created on that data sample. The thicker red lines show theoretical forms of model decay. The top plot shows the ideal situation in the top red line showing a continuously high level of model performance versus the realistic situation with the descending line of model performance. This plot is unrealistic due to data drift and model decay, which refer to the natural process of making less accurate predictions due to changes in data over time. In the bottom plot, the top red line shows a more realistic situation where the model is retrained each month, restoring the predictive accuracy to the maximum expected level each month. The overall gain is much greater from the frequently trained models than the originally trained model shown in the lower green line in each plot. However, this is a theoretical example based on data that was over-sampled to create multiple time periods.

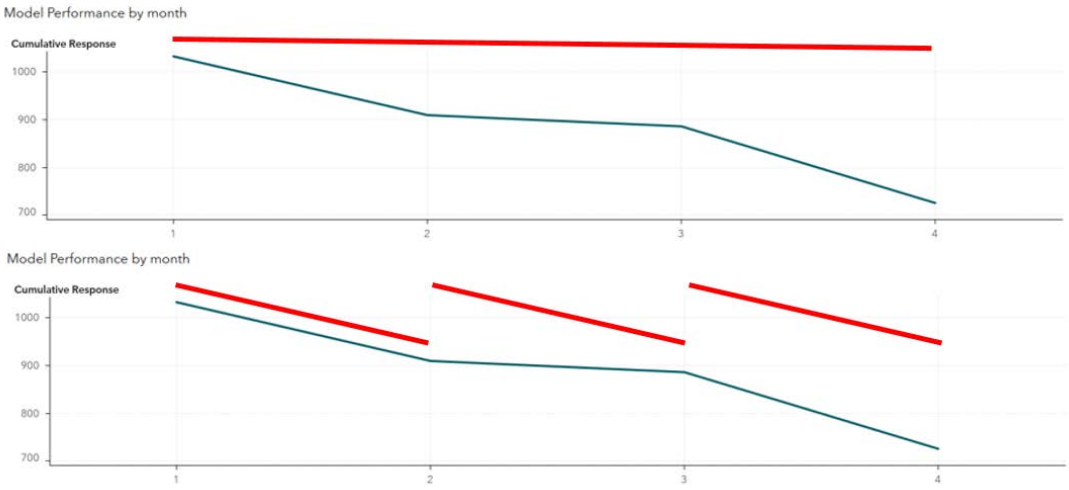

Figure 1. Theoretical Model Performance in Absence of Model Decay and by Frequent Retraining

MODEL MONITORING AND RETRAINING

There are numerous potential strategies for monitoring model performance and scheduling model retraining. Selection of a retraining strategy often depends on the business needs. Some processes can accept new models whenever they are created and validated. Some processes can accept new models only at fixed points in time. In many cases, models are created for comparison but never promoted to production. The main point is that model monitoring and retraining must be part of the business dynamic.

MONITORING

Model monitoring is a process for determining how well a model is or may be performing. There are several potential analyses that might be performed. Model monitoring process should measure all these factors:

- Data drift. Data values naturally changes over time due to numerous factors. People age. The economy becomes more or less positive. Mechanical parts erode or get updated. Competitors improve. Measuring changes in data values can be an early indicator of changes in model or business performance; however, not necessarily always.
- Model stability. Due to changes in data values, the distribution of model predictions may change. These changes will almost certainly impact business performance or planning. For instance, if predictions of truck maintenance-need increases, then more trucks will be scheduled for visits to the shop. More visits increase expenses regardless of the prediction accuracy.
- Model accuracy. If predictions target labels are available, then we may compute model accuracy measures. Degradation of model accuracy outside of acceptable bounds indicate a need for model retraining.
- Variable contribution. Changes in variable contribution to the model score or the model accuracy should be measured. These changes are also leading indicators of changes in model performance and may be used for reporting inferences about which variables caused changes in stability or accuracy. This may also be termed model interpretability.

The results of model monitoring should be stored and are used for model governance, statistical and business analysis, and as part of the process of determining if the model needs to be retrained.

RETRAINING

Model retraining is the process of recomputing a predictive or descriptive model on new data. Each new set of coefficients or effects is considered a new model. Models are retrained for multiple reasons.

- Business strategy. Changes to objectives such as increasing or decreasing acceptable levels of credit risk, investment in growth of new product lines, or numerous other facets will create the need for retraining models or creating new models.
- External conditions. Changes in business factors such as interest rates, new data sources, or suppliers of real-time truck metrics may create a need to retrain models.
- Business performance. Changes in measure such as response to promotions, credit repayment, truck repairs, and numerous others will create the need model retraining and / or review of the business strategy. Some change will be needed.
- Model Monitoring. Changes in the measures reported by model monitoring may create the need for retraining the model. This may be due to declining accuracy, data drift, or stability.

BUSINESS PROCESS

Organizations have many reasons for building predictive and descriptive models. Some models are used only for inference to learn more about the processes that shape the business or the expected impact of new strategies. Other models are created for integration into operational systems that interact with customer and business touchpoints to make the business more efficient, drive growth, improve loyalty, or other systematic objectives. The flow chart shown in Figure 2 is just one possible representation of a process for managing models.

The process flow is cyclical; however, we can say it starts with an operational business process that consumes and produces data. We are only representing the process for monitoring and retraining a model. We are not representing the process for defining a business problem and building the initial model. Here are the possible paths to start a model retraining process in this example:

- A timer event starts each cycle of the process, according to some predefined schedule.
- One timer event directly starts a model retraining. This is the process we are using in our simulation.
- Another timer event directly starts a model monitoring. The is the process we are using in our simulation.
- Another timer event checks for new data. If new data exists, a new monitoring job is executed. This could also be the process we are using in our simulation. New flight data arrives in monthly chunks.
- Regardless of the source, we always want the monitor process to record the current statistics for future analysis.

- A KPI measures computed from the monitoring output may drive the retraining. For instance, we may want to retrain if model accuracy falls below a threshold such as misclassification greater than 20%.
- A business strategy change may trigger a model retraining if not a completely new model.
- The newly retrained model should be tested for measures of robustness, accuracy, or expected ROI. It may be compared to a champion model. The model may fail testing and trigger a review of the model building process.
- If a new model passes testing, it may be deployed into the production environment for integration into the operational business process.

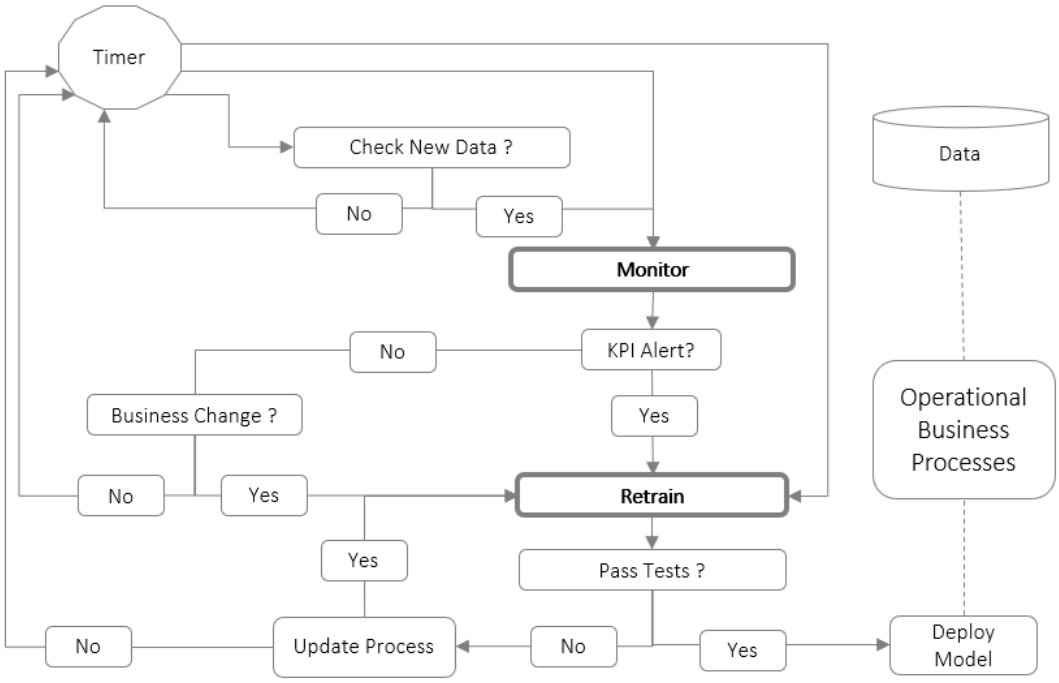

Figure 2. Sample Process Flow Diagram for Model Monitoring and Retraining

IMPLEMENTATION

The goal of our analysis is to test the effect of different strategies for model monitoring and retraining on long term model performance. To create this very custom process, completely new SAS code was written. Here are the descriptions for the major components of the code:

- The primary data was downloaded from the Bureau of Labor Statistics web site. The data consists of 145M rows of data stored in multiple CSV files.
- PROC IMPORT was used to import each CSV file into a corresponding SAS data set. Minimal data cleaning was performed at this stage. Several variables that are naturally numerical integers were mistakenly imported as character variables in the SAS tables and needed to be changed in the next step.
- DATA step and Base SAS procedures were used to transform and clean the data. A small number of observations had missing values for departure or arrival time and

were removed from the data. Several character variables were transformed into numerical columns. Variables that were irrelevant to the analysis were dropped. Variables that were proxies for flight late arrival were dropped. The target variable *LATE* was created with numerical Boolean value of (*Arrival_Delay > 15).* All months of data were combined into one large table with 145M observations. This "big table" was for all calculations.

- DATA step was used to create Training and Monitoring samples by querying the big table for specific months of data. The Training data was divided randomly into approximately equal samples of Train and Test data. Train data was used to build the model. Test data was used to report the statistics from the training exercise.
- SAS High-Performance Analytics procedures were used to create Decision Tree and Logistic Regression models. Default settings were used in all cases. Score code was saved from each training run into a directory of files. The score code was used to compute test data statistics and for model monitoring. PROC HPSPLIT and ODS were used to create the Decision Tree display images.
- Base SAS procedures were used to test statistics and model monitoring statistics such as mean monthly values of *Late* proportion, Probability, Misclassification, and True Positive rates.
- PROC SGPLOT and PROC PRINT were used to make all graphs and table displays.
- The SAS macro %SIM was used to script these operations. The %SIM macro was developed to simulate model retraining and monitoring with different time periods for the entire 303 months of data. All statistics used in this paper came from the %SIM macro.

Note: All SAS code that was used for this paper is available from the author upon request.

DATA

For the remainder of this paper, we will refer to the Airline flight data used in several data mining competitions and samples. The data is freely available from the U.S. Bureau of Transportation Statistics. The data starts in October of 1987 and continues to be updated. Our sample ranges from 1987 until the end of 2012. The data contains variables describing various attributes about the flight including the scheduled arrival time and the actual arrival time. Several papers have been written about this data including a visualization paper by Rick Wicklen as contribution to the ASA Data Expo contest in 2009.

We use this data because it represents a consistent source of data over many years, which has the potential to show change in data values and patterns over that time. In this exercise, our goal is to show long term trends in model monitoring; we are not trying to infer new knowledge about the data or build the most sophisticated model. Our sample contains 145,664,836 observations. All variables that would not be available at the time of model building or model deployment have been rejected. The first ten rows of data are printed in Table 1.

Obs	late	ORIGIN	DEST	UNIQUE_CARRIER	DAY_OF_WEEK	DAY_OF_MONTH	MONTH	YEAR	DISTANCE	CRS_DEP_TIME	CRS_ARR_TIME	CRS_ELAPSED_TIME	DEP_DELAY
1	0	JFK	LAX	AA	4	1	10	1987	2475	900	1152	352	1
2	0	LAX	HNL	AA	4	1	10	1987	2556	1300	1535	335	4
3	0	LAX	JFK	AA	4	1	10	1987	2475	830	1640	310	8
4	0	OGG	HNL	AA	4	1	10	1987	100	2035	2110	35	3
5	0	JFK	LAX	AA	4	1	10	1987	2475	1200	1446	346	2
6	0	DFW	HNL	AA	4	1	10	1987	3785	945	1250	485	2
7	0	HNL	OGG	AA	4	1	10	1987	100	1345	1430	45	-2
8	1	IAH	DFW	AA	4	1	10	1987	224	737	843	66	0
9	0	DFW	STL	AA	4	1	10	1987	550	825	1010	105	1
10	0	HNL	DFW	AA	4	1	10	1987	3785	1929	735	426	0

Table 1. Sample of Data Showing Variables with Typical Values

The derived target variable is named *late* and is either 0 or 1 to indicate more than 15 minutes late. The variables starting with CRS are scheduled times. The only variable that depends on the instance of the flight is departure delay, *DEP_DELAY*, *w*hich is necessary to produce good models without creating complicated lag variables.

The business of managing flight on-time performance has many latent factors. Airlines are reported to implement procedures to control and improve their on-time percentage as needed. They may use this data to make announcements about their performance and enhance their marketing campaigns. Flights that leave late may spend more fuel in an effort to regain time. Flight crew and airport expenses may constrain on time performance.

Table 2 shows the number of flights aggregated by month over the entire time period. Column N refers to the total number of flights. The monthly late rate averages 19.0% and ranges from 10.2% to 32.0%. Numeric model input predictor variables are also shown. The scheduled elapsed time, *CRS_ELAPSED_TIME*, shows a notably small standard deviation, perhaps indicating there has been little overall change in the scheduled routes.

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
late	145664836	0.1904100	0.0443149	0.1029323	0.3204210
DISTANCE	145664836	710.7584869	46.5993466	587.8776336	788.1398129
CRS_DEP_TIME	145664836	1332.88	26.8047265	1173.91	1361.85
CRS_ARR_TIME	145664836	1492.18	30.0217495	1303.90	1518.86
CRS_ELAPSED_TIME	145664836	122.9232413	8.2599934	99.5262048	136.6999134
DEP_DELAY	145664836	8.1111819	2.9086702	2.2205164	17.1477836

Table 2. Aggregated Monthly Means for the Entire Period of 303 Months

The plot of the number of flights per month is more interesting, in Figure 3. The sample contains 303 months of data over 25 years. The small yearly seasonality is apparent. There are peaks in travel around the winter holidays and over the northern hemisphere summer vacation periods. Markers have been added for selected significant global events. The dramatic impact that the September 11, 2001 terror attacks had an obvious impact on air travel, as expected, followed by a dramatic rise in the number of flights in January 2003. The rate of flights that are late each month is shown in the lower plot. There is minimal correlation between the total number of flights and the rate of late flights.

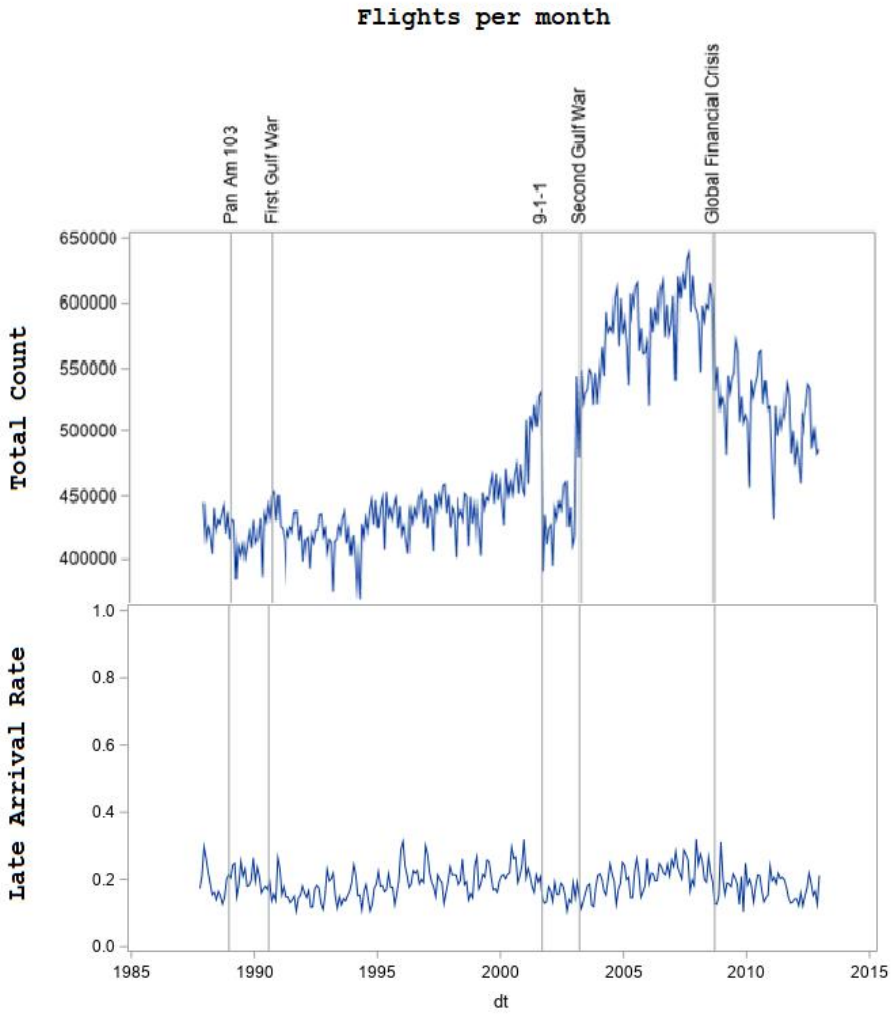

Figure 3. Total Number of Airline Flights per Month with Significant Event Markers

Most treatments of this data focus on modeling or visualizing the entire data set. However, imagine that you are an analyst working in 1987.

MODELS

The data is provided in monthly data sets. We created our first model on the first month of data, October of 1987, which contains 448620 rows. The data is randomly split into half training data and half test data. The model is a default decision tree created by PROC HPSPLIT, which uses 10-fold cross validation to control the growth of the tree. A decision tree is good default model for this study since it is tolerant to new data values and naturally incorporates variable selection. Figure 4 illustrates the model results with the complete classification tree for the first month of data and the top subtree with details about the variables used in the model.

The complete classification tree demonstrates a complex model using several variables. The categorical variables identifying the airline, origination airport, and destination airport have higher cardinality and contribute to many of the tree branches. The subtree view shows top portion of the tree where departure delay is the most significant variable, as expected, but that other variables contribute to the classification values.

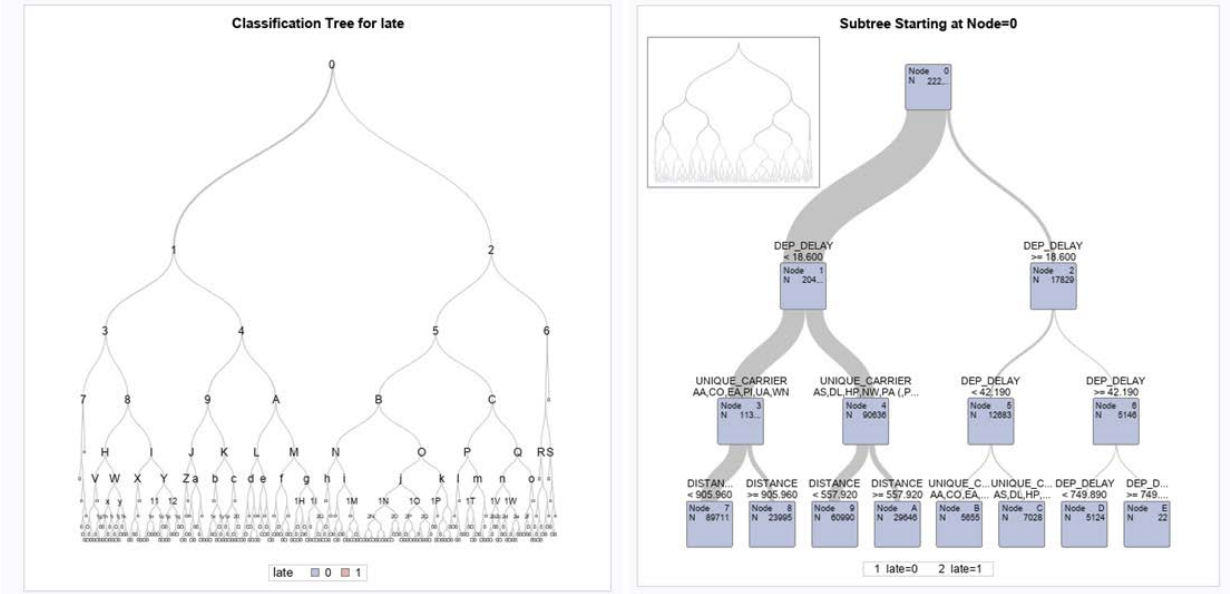

Figure 4. Classification Tree and Subtree of Model Variables

The relative variable importance values are shown in Table 3. These values correlate with the detail view of the decision tree. After *DEP_DELAY*, the remaining variables retain significant impact on the classification rates.

Variable Importance			
	Training		
Variable	Relative	Importance	Count
DEP_DELAY	1.0000	143.6	4
UNIQUE_CARRIER	0.1955	28.0687	18
DAY_OF_MONTH	0.1588	22.8046	71
DISTANCE	0.1517	21.7843	20
CRS_ELAPSED_TIME	0.1040	14.9300	28
ORIGIN	0.0916	13.1512	24
DEST	0.0907	13.0277	37
CRS_DEP_TIME	0.0898	12.8917	29
CRS_ARR_TIME	0.0831	11.9399	20
DAY_OF_WEEK	0.0285	4.0969	6

Table 3. Decision Tree Variable Importance Measures

Table 4 demonstrates the Decision Tree model results on the test sample from the first month of data. Late is the proportion of late flights, *i_late* is the proportion of flights classified as late, *i_misc* is the overall misclassification rate, and *i_tp* is the proportion of flights correctly classified as late. The score code was then applied to the test data sample and the classification (*i_late*), misclassification flag (*i_misc*), and true positive flag (*i_tp*) were computed. PROC MEANS was run to summarize the test scores and produced the data shown in Table 4. We can see that in the test sample, 17.3% of the flights were late, 7.2%

of flights were misclassified, and 10% of flights were correctly classified as late. These metrics will be used to monitor model accuracy in the remaining data.

First month model

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
late		222609	0.1727199	0.3780058	0	1.0000000
P_late0	Predicted: late=0	222609	0.8273965	0.2284448	0	1.0000000
P_late1	Predicted: late=1	222609	0.1726035	0.2284448	0	1.0000000
i_late		222609	0.0809446	0.2727507	0	1.0000000
i_misc		222609	0.0721220	0.2586903	0	1.0000000
i_tp		222609	0.1005979	0.3007962	0	1.0000000

Table 4. Decision Tree Model Results on Test Sample from the First Month of Data

For comparison purposes, we also ran a Logistic Regression model through the same process. The absolute results are similar as shown in table 7. The misclassification rate is 1.2% higher, and the true positive detection is 1.5% lower.

First month model -- Logistic Regression

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
late		222609	0.1727199	0.3780058	0	1.0000000
P_late0	Predicted: late=0	222609	0.8277114	0.2414860	6.11583E-102	1.0000000
P_late1	Predicted: late=1	222609	0.1722886	0.2414860	0	1.0000000
i_late		222609	0.1029249	0.3038614	0	1.0000000
i_misc		222609	0.0878311	0.2830498	0	1.0000000
i_tp		222609	0.0848888	0.2787167	0	1.0000000

Table 5. Logistic Regression Model Results on Test Sample on the First Month of Data

The results are not as good as the Decision Tree. Table 6 was generated to compare the two models. The table illustrates the comparison of models on first month on test data. Variable *ms* is the sequential month counter. TPR is the sample true positive rate. Decision Tree is champion based on misclassification and true positive rates. We can conclude that we have a valid modeling process using the default decision tree and will use that for the following results.

First month models - comparison

model	_FREQ_	ms	late	i_late	i_misc	i_tp	TPR
Decision Tree	222609	1	0.17272	0.08089	0.072279	0.10044	0.58152
Logistic Regression	222609	1	0.17272	0.10292	0.087831	0.08489	0.49148

Table 6. Comparison of Models on First Month on Test Data

It would have been tempting to build models on the entire data that account for all the seasonality, long term trends, forecasts, and significant events ahead of time. However, we must put ourselves into the position of the analyst in November of 1987 who received this minimal data with the task of producing models that would give the best prediction for each flight as it happens. We would start with only one month of data.

A good question to ask is what decision tree models would have been available to the analyst in 1987, and what kind of computers would have been used. Brieman et al. published "Classification and Regression Trees" in 1984 and Quinlan published "Decision Trees as Probabilistic Classifiers" in 1987. For our purposes, the general answer is good enough. We can proceed using Decision Trees. However, we should consider that in a real life situation we should evaluate new models at every opportunity for improving a model retraining process.

SIMULATIONS

The next task is to see how that model performs on subsequent months. We built our first model on a training sample from October. We then scored the model on all the data from October, November, and December, giving us three full months of history.

THREE-MONTH RESULTS

We come back to work in January to see how we are doing. Figure 5 illustrates the plots of the monthly proportion of flights that are late, misclassified, and correctly classified as late. The first three months of model monitoring show decreasing rates of accuracy and the scatter plot shows a possible relationship between accuracy and proportion late. The plots are unspectacular but appear to show trends. The proportion of late flights and the misclassified rate are increasing. The true positive rate is decreasing. This appears to follow the theory perfectly, as data changes over time that model accuracy and performance degrades.

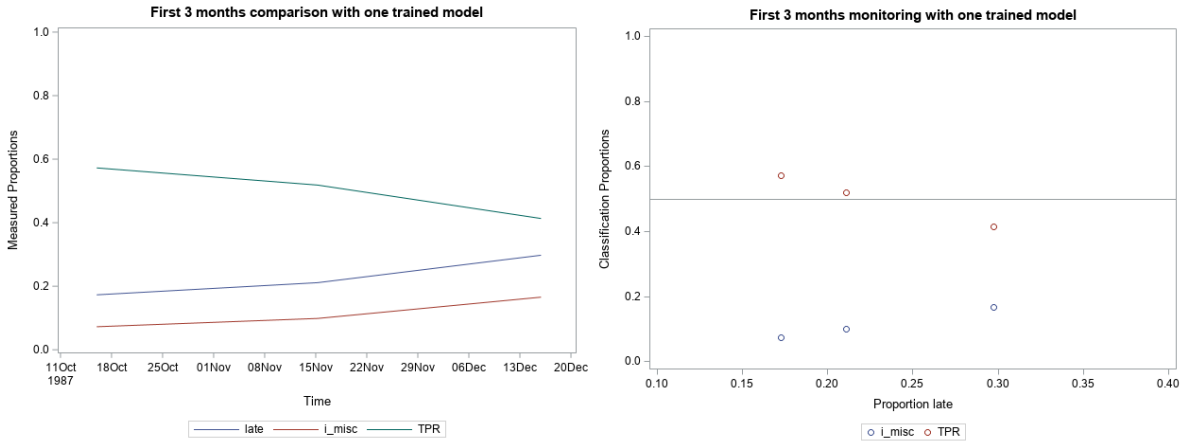

Figure 5. First Three Months of Model Monitoring and Proportion Relationship

We now have a decision to make. Should we wait another three months to see what happens? Or should we build a new model now and risk overfitting a short-term trend? We decide to do both. We will build a new model and compare the two strategies after we take a vacation and return in April.

SIX-MONTH RESULTS

We come back from our ski vacation in April 1988 to examine the results. First, we look at the results from the single model we trained based on data from October. The surprising results are shown in Figure 6. After the model decay observed in December, the model performed more accurately in months January, February, and March. This correlation with the changing proportion of late flights is marked. We can hypothesize that the pattern of late flights is different for the very busy month of December.

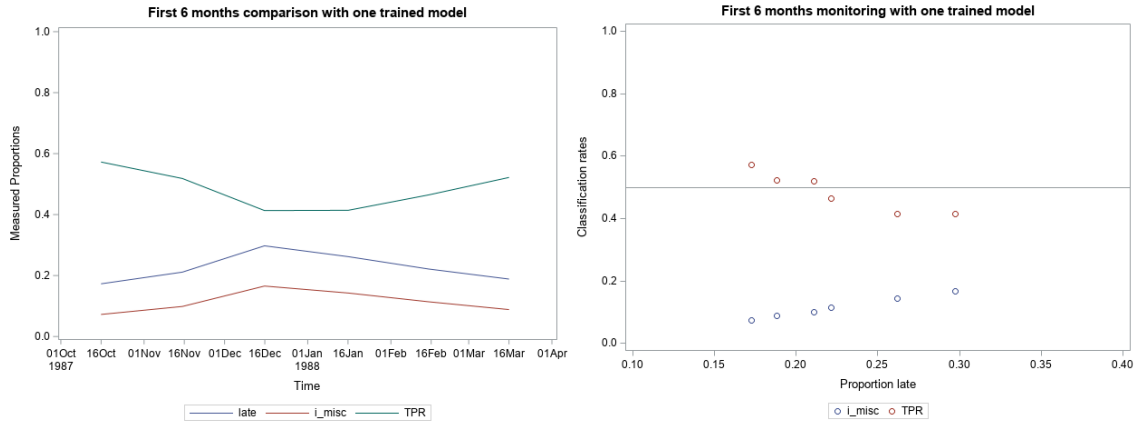

Figure 6. Six-Month Results on for the Model Originally Trained on using October Data

These scattered results are not definitive. To find a better answer, we trained and monitored models across all combinations of time periods. This included multiple-month training periods and multiple-month monitor periods for the first six months. We sampled 24 different combinations. The resulting matrix of data was fed into PROC SGPLOT to create the heatmap shown in Figure 7. The color response statistic is the mean. The most accuracy monitor periods have the most training time in months. The best continuous solution across the sample is the diagonal where the most possible training months were used to create the model monitored in the subsequent month. The best discontinuous solution is the diagonal up to month 6 when three or four months of training data were better than five.

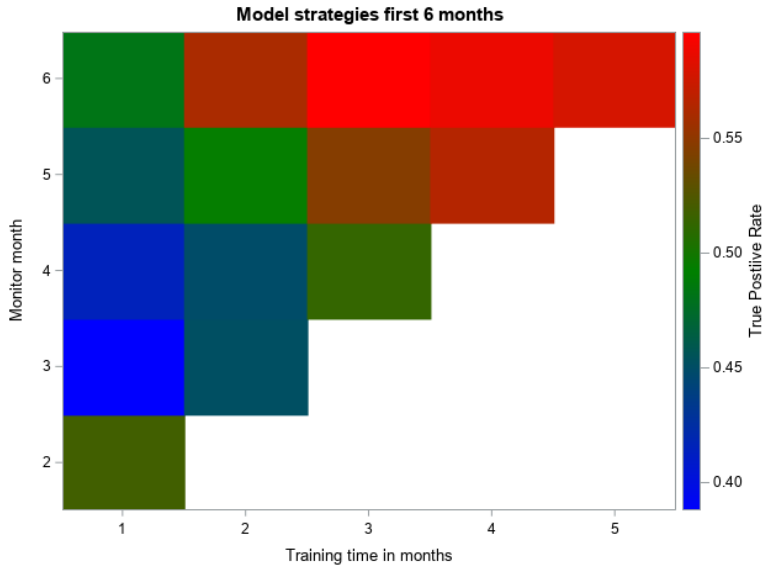

Figure 7. Model Monitoring by Training Time Period in Months

ONE YEAR RESULTS

Based on the knowledge we gained in April that longer training periods performed better on future data, we tested four strategies for the remainder of the first twelve-month period. The standard naïve single model and a model trained on each month of data are shown in Figure 8. We focus on the true positive rate as that measure that will most impact our ability to identify and react to flights that are predicted to be late. In Figure 8, the monitored true positive rate is below 50% in most months. It is surprising that the single model trained using October data is a better predictor of the next 12 months than the set of eleven models trained to predict only one month ahead for the next 12 months. However, neither model strategy is promising.

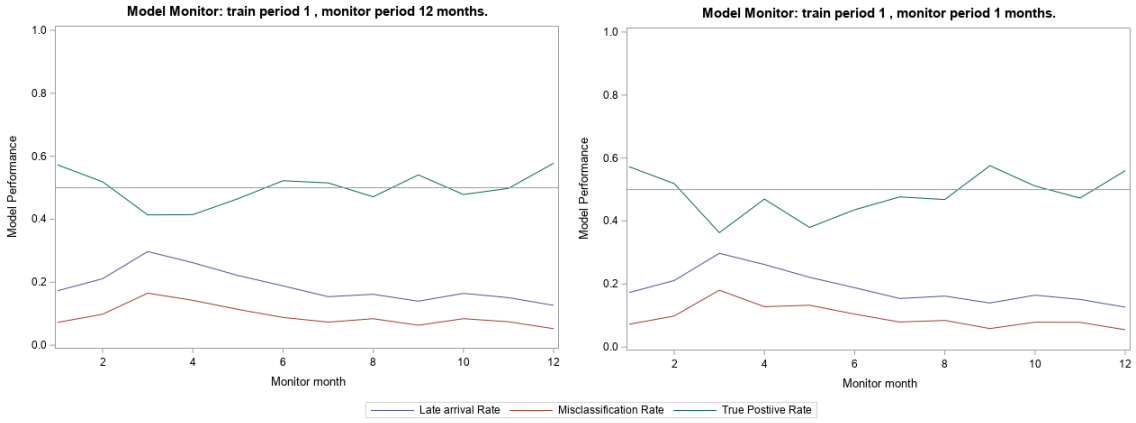

Figure 8. Model Performance of Naïve Models Trained on One Month and Each Month

Our next strategy is to test long model training periods. Each of these strategies improve performance as displayed in Figure 9. They show a much-improved true positive rate over the naïve models with true positive rates greater than 50% most of the time. In particular,

the model based on four months of training time and three months of monitoring time did very well. This is likely due to including enough data to capture periodic effects. The data is known to have seasonal patterns. There are more flights around the winter holidays and summer vacations. There are also more weather delays in the northern hemisphere in winter. However, at this point in time, October 1998, we do not have enough data to conclude that periodicity is a main effect.

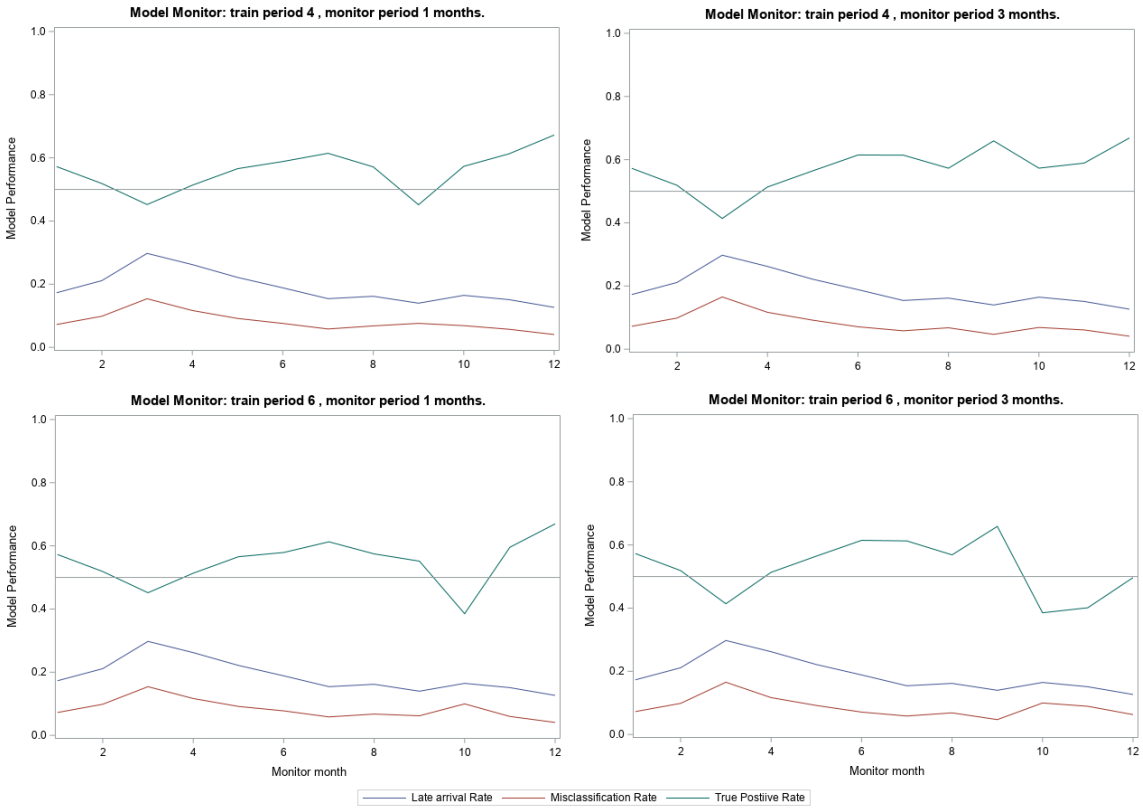

Figure 9. Four Model Training and Monitoring Scenarios with Longer Training Periods

Based on these results, we will apply the 4-3 model (4 training months to 3 monitor months) to the remainder of the data. Every three months a new model will be created using the previous four months of training data. This creates a one-month overlap in training data between consecutive models, which helps smooth the changes from one model to the next.

This strategy will result in 100 new model training events. Since we are creating both a decision tree model and a logistic regression model, that will create 200 models. Each model will be created on four months of data. Months have on average approximately 500 thousand observations and we use half the sample for training and half for testing thus resulting in training samples of approximately 500K*4/2= 1 million rows, depending on the actual airline traffic for those months. The total amount of data used in training models will be approximately 200*500 M = 10000 M = 10 billion rows!

TWENTY-FIVE YEAR RESULTS

Now jump to the beginning of the year 2013. It has been 25 years and three months since we started this project. We have been building models and monitoring their progress during that time. Before we retire from our cushy data scientist position, we will take one more look at the relative model performance of each strategy.

The single model (1-303) strategy now produces an expected result. The green top line is the true positive rate, TPR, which shows a downward trend in expected value. The blue middle line is the actual proportion of late flights, which does not show a strong long-term trend. The bottom red line is the monthly misclassification rate that shows a slight upward trend. However, we don't yet understand the pattern changes that cause this decay in performance; that work is outside the scope of this paper.

We have also been running the four-three strategy where each model was trained on the four most recent months of historical data and then monitored for the next three months. The difference is not as great as expected based on the first year of performance. The baseline model strategy has a mean monthly TPR of 0.382; the four-three strategy produces 0.411. Neither strategy is compelling.

Since we now have 25 years of data, we can test other long-term strategies. We believe that there are seasonal effects from monthly up to yearly if not longer. Therefore, we tested additional strategies training data on twelve and eighteen months of history. The 12-6 and 18-6 results simulations produce incremental improvements. The results are plotted in Figure 10 and listed in Table 7. Each strategy produces different cycles of better and worse model performance, and all show levels of the long-term trend to worse TPR values. Further studies might discover a strategy or model function that produces better and more reliable results.

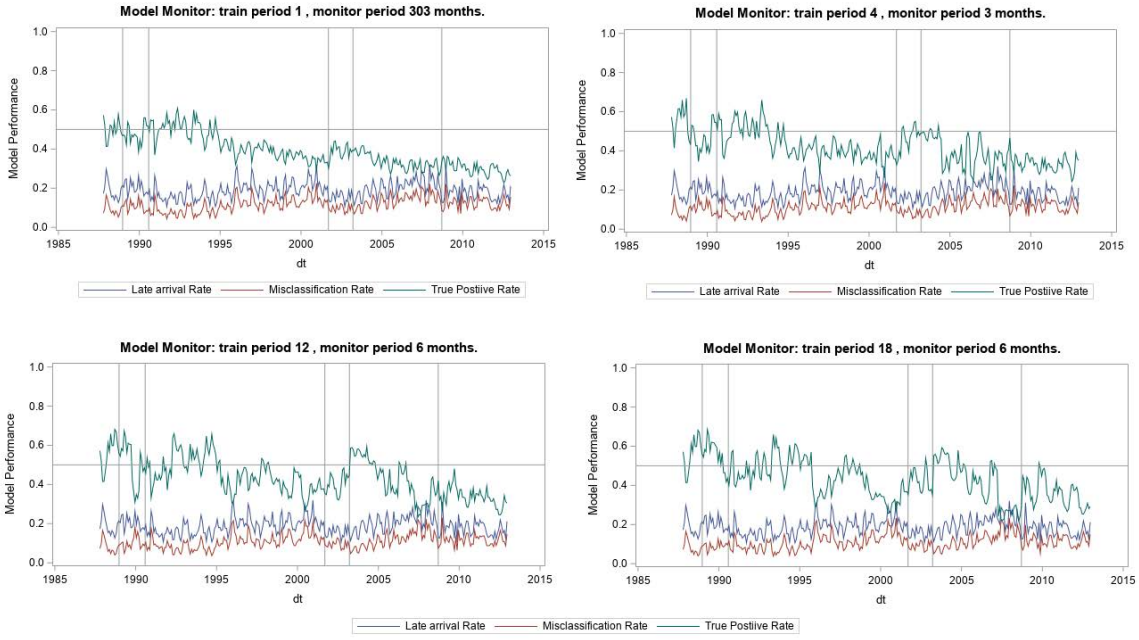

Figure 10. Long Term Model Performance of Multiple Retraining Strategies

The final column of Table 9 is the rate of months that have a true positive rate greater than 0.5. This could be an important measure of model usefulness. None of the scenarios reliably produced models with a monthly TPR greater than 0.5. This shows a weakness with

all the models used in this exercise. The last two cells are highlighted as they show a significantly elevated monthly TPR. The training fit statistics, computed on test data, indicate we may have hit the limit on core model accuracy.

Strategy	Training			Monitoring: 303 months			
	Months per model	Mean Misclassification Rate	Mean TPR	Months per model	Mean Monthly Misclassification Rate	Mean Monthly TPR	Monthly TPR > 50%
Baseline	1	0.072	0.581	303	0.117	0.399	0.134
4-3	4	0.119	0.400	3	0.113	0.411	0.155
12-12	12	0.135	0.411	12	0.111	0.422	0.207
12-6	12	0.112	0.418	6	0.110	0.428	0.249
18-6	18	0.113	0.415	6	0.111	0.427	0.270

Table 7. Comparison of Model Retraining Strategies

CORRELATION

Another aspect is correlation between model performance and the proportion of late flights as displayed clearly in Figure 11. The plots of the baseline 1-303 strategy and the best-performing 18-6 strategy are shown in Figure 11. In both cases, the misclassifications correlate well with the target variable, but the true positive rate shows significant dispersion. The 18-6 models show more true positive values above 50% especially across the greater vales of late arrival rate. This gives us more confidence in the 18-6 strategy.

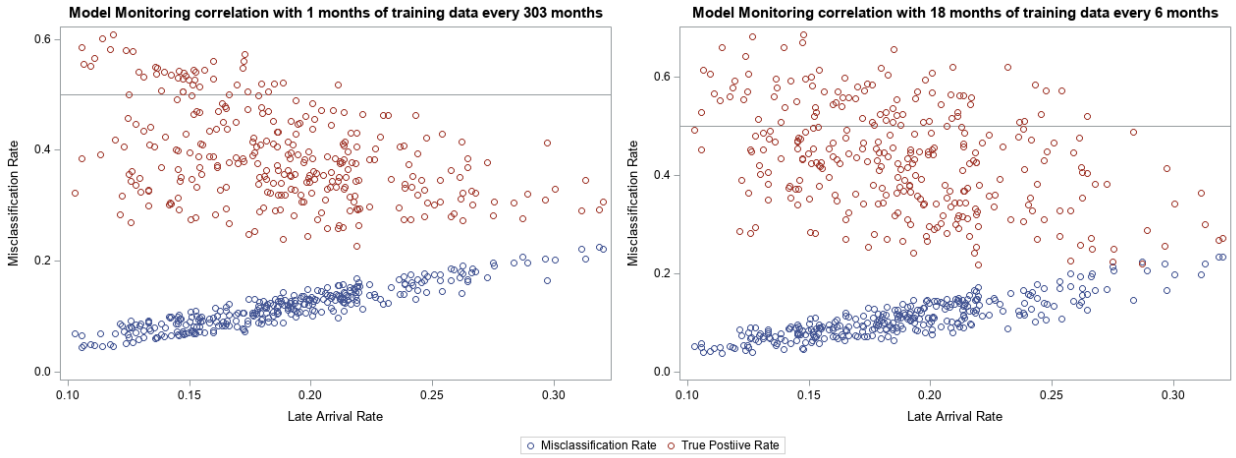

Figure 11. Long Term Correlation between Classification Measures and Late Proportion

SAS MODEL MANAGER

The SAS® Model Manager product contains many of the capabilities shown in this paper. You can register all models that were creating in this exercise into a versioned repository by using a GUI application, SAS macros, Python code, or REST API services. You can execute model monitoring tasks that are similar to the ones presented here with additional capabilities. These capabilities include computing variable distributions, input and output

variable drift, and rank order statistics such as lift, captured response, KS, and Gini. SAS Model Manager can also compute a Feature Contribute Index to measure the correlation between predicted values and input variables over time. SAS Model Manager provides workflow capabilities to manage the business process shown in Figure 2. Finally, SAS Model Manager can test and deploy SAS and Python models to both batch and real-time servers for operational integration. However, SAS Model Manager does not have the extensive simulation capabilities show in this paper. Most users are expected to be working in the moment, rather than analyzing twenty-five years of data. If you are interested in this capability, contact the author for more details.

CONCLUSION

Model monitoring and retraining are key parts of any operational model scoring process. Many paths can lead to model retraining. In this work we studied retraining models at regular intervals over a very long running process that has produced 25 years of data. The length of the time period of data used to train the models and the length of time monitoring the models in production have significant impacts on lifetime model accuracy. Data scientists should carefully monitor their models and conduct experiments to optimize those parameters.

The simulation capabilities developed for this paper were useful in testing different combinations of retraining and monitoring parameters. We found that this data contains both short and long-term periodic effects. The best combination of parameters we found used an 18-month sample to predict a 6-month interval. The core finding is that a training period should be long enough to accommodate periodic effects and should be longer than the monitoring period. We cannot generalize that specific recommendation to every process, but we want to highlight the need for observing and adjusting model retraining and monitoring. The simulation framework could be extended to test additional parameters and scenarios.

The Airline On-Time flight data from the National Bureau of Transportation Statistics continues to provide a rich source of publicly available data. The data is now complete from the 1987 through 2019.

Future work could go in several directions. We should study the effects of implementing champion-challenger strategies and dynamically changing the champion model as accuracy decreases. We should study the possibility of using forecasting to estimate when models might need retraining especially in the presence of seasonal or long-term effects. We should look at using optimization to dynamically adjust the training and monitoring parameters.

A final key finding is that the SAS system makes a great platform for importing and cleaning extremely large amounts of data, and for computationally processing that data over long periods of time. Each simulation processed billions of records over hundreds of iterations within several hours. At the end the same software was able to summarize the results and produce useful and professional tables and graphs.

REFERENCES

Breiman, Leo, Jerome Friedman, Charles J. Stone, R.A. Olshen. 1984. *Classification and Regression Trees*. Taylor & Francis

Quinlan, J.R., 1987. "Decision Trees as Probabilistic Classifiers." *Proceedings of the Fourth International Workshop on Machine Learning*. Pages 31-37. University of California, Irvine.

Wicklin, Rick. 2009. "An Analysis of Airline Delays with SAS/IML® Studio." *Proceedings of the ASA Data Expo 2009*. Cary, NC: SAS Institute Inc. Available https://support.sas.com/rnd/app/iml/papers/abstracts/airlinedelays.html.

Bureau of Transportation Statistics. 2019. "Database Name: Airline On-Time Performance Data." Available https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time (accessed November 2019).

Duling, David. 2019. "The Aftermath What Happens After You Deploy Your Models and Decisions." *Proceedings of the SAS Global Forum 2019 Conference*. Cary, NC: SAS Institute Inc. Available https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3496-2019.pdf.

ACKNOWLEDGMENTS

The author wishes to thank many people who make and support SAS software. In addition, the author thanks both Ming-Long Lam and Phil Easterling, both of SAS Institute, for their advice on both data mining and the airline industry, and Kristen Aponte for prompt editing.

RECOMMENDED READING

- *SAS Model Manager 15.3 Users Guide*
- *SAS/STAT 15.1 User's Guide: High-Performance Procedures*

CONTACT INFORMATION

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

David R Duling
SAS Institute Inc.
919 793 5663
David.Duling@SAS.COM