SAS® GLOBAL FORUM 2020 #SASGF

Paper 4693-2020

Winning Tactics for SAS® Visual Data Mining and Machine **Learning: Why it's the Only Skill You Need**

Chris St. Jeor, Zencos

ABSTRACT

The ability to rapidly identify actionable data insights determines today's business success. SAS® Visual Data Mining and Machine Learning on SAS® Viya® is a one-stop shop that enables both technical and non-technical stakeholders to complete that task with ease. From data exploration to machine learning, SAS Visual Data Mining and Machine Learning enables analysts to load data and find the needed answers quickly and efficiently.

In this workshop, users will gain hands-on experience using SAS Visual Data Mining and Machine Learning as you walk through an analytics project from start to finish. Users will learn how to import and explore the data, build several predictive models, and then determine a champion model to answer business questions. Users will walk away with a baseline knowledge of the application and the various control mechanisms it provides.

For a fun twist on common everyday business problems, this workshop uses college basketball data to predict the outcomes for tournament games.

INTRODUCTION

The average machine learning or predictive modeling project is a time-consuming and painstaking process. The user must clean the data, explore variable relationships, code and recode and recode and recode predictive models, publish the champion model, and then reevaluate the model every few months. This process typically requires several software platforms, packages, and skill sets. The process becomes more complex when the user or users need to collaborate on a project. SAS Visual Data Mining and Machine Learning was designed to help mitigate these challenges. Visual Analytics and Machine Learning was specifically created to allow users to work through an end-to-end analytics process in one central location without having to type a single line of code.

Learn how SAS Visual Data Mining and Machine Learning on SAS Viya® can orchestrate the compete analytic life cycle within one platform, from preparing the user's data through governing the user's models – all with just a few clicks of the mouse. While SAS has a strong reputation as a programming language, Visual Data Mining and Machine Learning allows users to find powerful insights with no coding necessary.

This paper will provide a baseline introduction of how to use SAS Visual Data Mining and Machine Learning. It will walk through step-by-step directions of how to create a predictive model and register the model for future use. Readers will specifically learn how to perform the following tasks:

- Create a new project
- Import and partition a modeling dataset
- Explore variable relationships
- Build a baseline model
- Create an analytics pipeline

- Evaluate model performance
- Register and save a champion model

The examples in this paper use college basketball data to predict which teams will win in the NCAA Men's Basketball Tournament.

SAS DRIVE

When a user signs into SAS Visual Data Mining and Machine Learning they will first see an interface called SAS Drive. SAS Drive essentially serves as the analytics homepage where users can create new content, navigate to previously saved projects, or access projects that have been shared with them. The ability to access shared projects should not be overlooked as it allows easy collaboration across teams in one shared space.

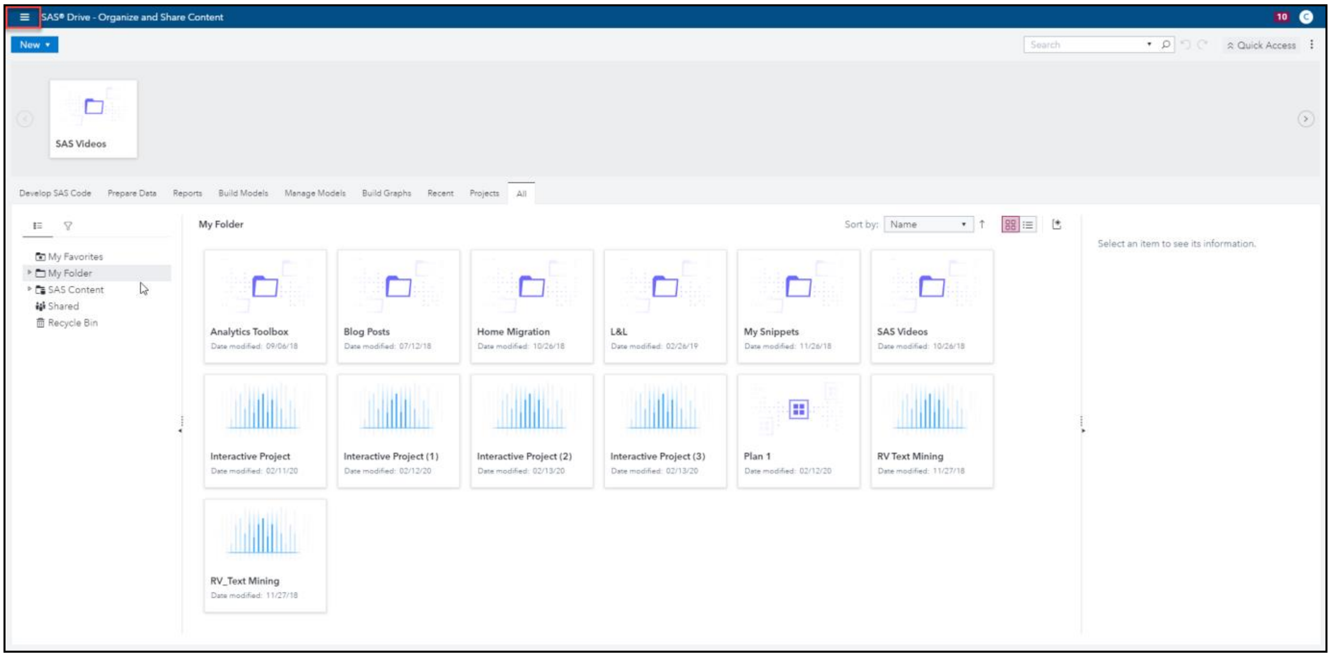


Figure 1. SAS Drive Interface

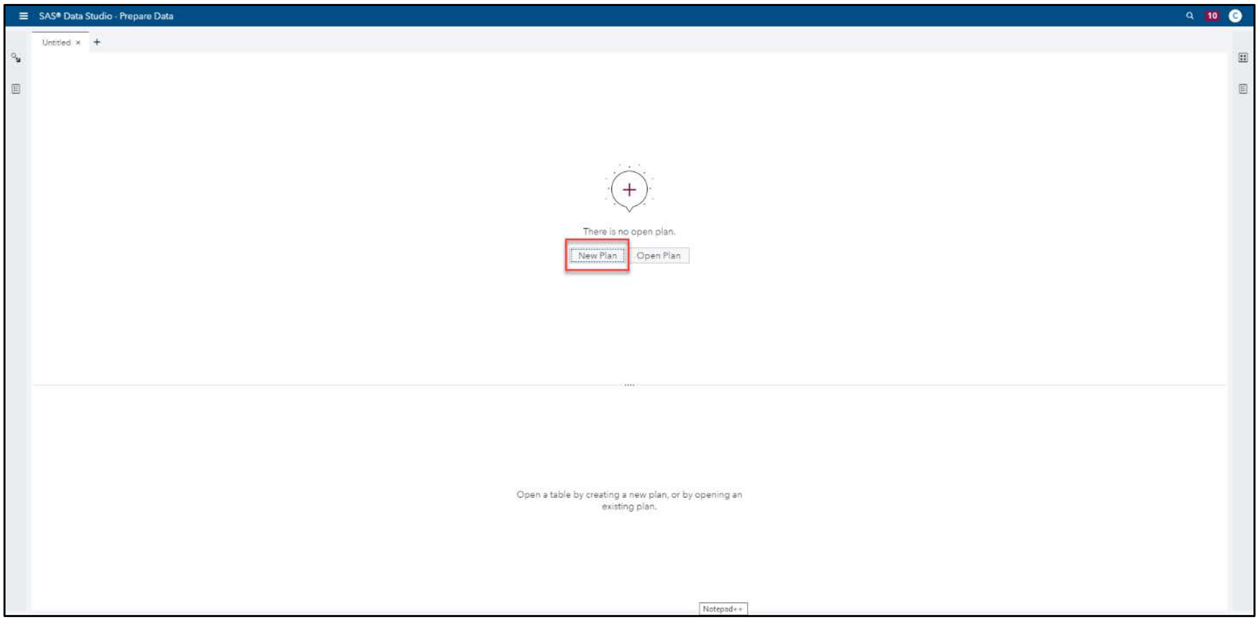
Clicking on the hamburger menu at the top left of the page will expand the full list of capabilities that Visual Data Mining and Machine Learning provides. Users can manage and prepare data, visualize data, build models, and more, all in the same application.

Visual Data Mining and Machine Learning is a fully unified platform, meaning that users can navigate through the entire analytics process without ever leaving the current browser tab.

For the purpose of this exercise you are going to start with a new project. The first step is to load and prepare the data. Click on the hamburger menu at the top left of the page and choose on the Prepare Data option. This will take us to SAS Data Studio where you can upload the data and do any necessary data cleaning and partitioning.

DATA STUDIO

SAS Data Studio, as seen in the figure below, is a user interface that allows users to import data, clean data, filter data, manage variables, etc. First, the user should create the dataset for the modeling project. Because a new project is being created from scratch, users must click on the "New Plan" button, highlighted in red.



There is no open plan.

Open a table by creating a new plan, or by opening an existing plan.

Figure 2: Data Studio

Once this icon is clicked, a new pop-up appears, shown in the figure below. Any data sources that have previously been loaded to the environment will be made available in the Available and Data Sources tabs seen below. By default, the Import tab is selected, which is what should be used in order to read in the locally saved data. Multiple methods are available to load data for the first time. The data file can be dragged and dropped onto the screen, or the user can navigate to the data by clicking the Local File icon highlighted below. Clicking this will open a new window to navigate to the locally saved dataset. Once the desired data is selected, click Open.

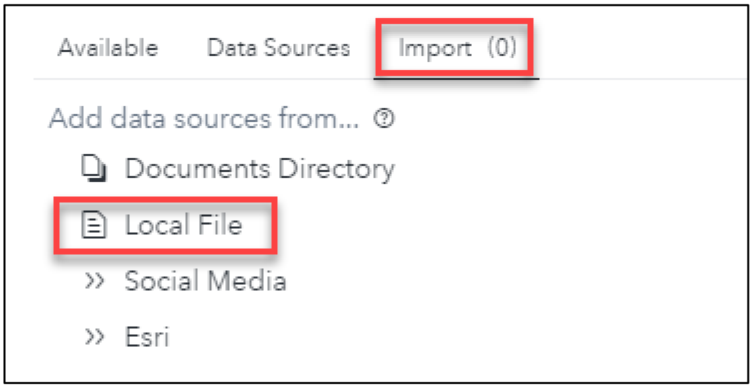
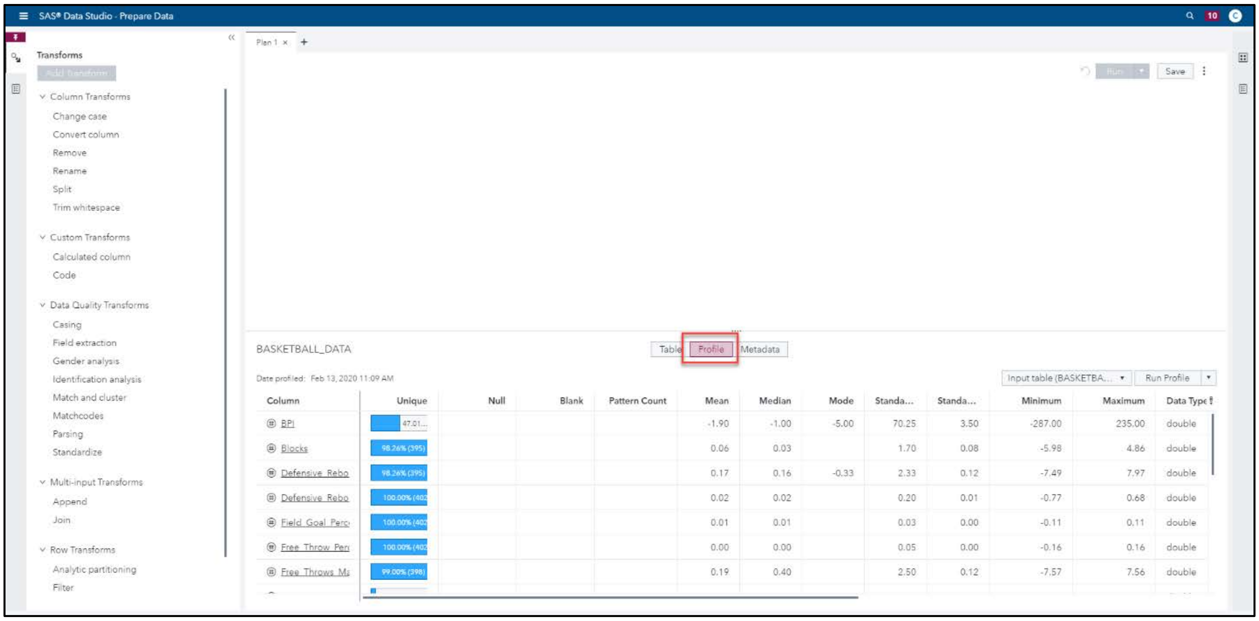


Figure 3: Data Import

Now that the data is loaded, another window will load that provides various options such as creating a required password or encryption key to prevent other users from access the data. Once the appropriate protections are in place, simply click the Import button and then OK.

The data is now loaded into memory on the CAS server and the user has several options available in order to make any additional edits or transformations required. A good starting point with any project is to make sure the data loaded correctly. On the main working window, click the "Profile" tab highlighted below. This is the equivalent of running proc freq and proc means on the dataset. The output can be seen below.



BASKETBALL_DATA

Date profiled: Feb 13, 2020 11:09 AM

Column	Unique	Null	Blank	Pattern Count	Mean	Median	Mode	Standa...	Standa...	Minimum	Maximum	Data Type
BPI	47.01...				-1.90	-1.00	-5.00	70.25	3.50	-287.00	235.00	double
Blocks	98.26% (395)				0.06	0.03		1.70	0.08	-5.98	4.86	double
Defensive_Rebo	98.26% (395)				0.17	0.16	-0.33	2.33	0.12	-7.49	7.97	double
Defensive_Rebo	100.00% (402				0.02	0.02		0.20	0.01	-0.77	0.68	double
Field_Goal_Perc	100.00% (402				0.01	0.01		0.03	0.00	-0.11	0.11	double
Free_Throw_Perc	100.00% (402				0.00	0.00		0.05	0.00	-0.16	0.16	double
Free_Throws_Ma	99.00% (398)				0.19	0.40		2.50	0.12	-7.57	7.56	double

Figure 4: Data Profile

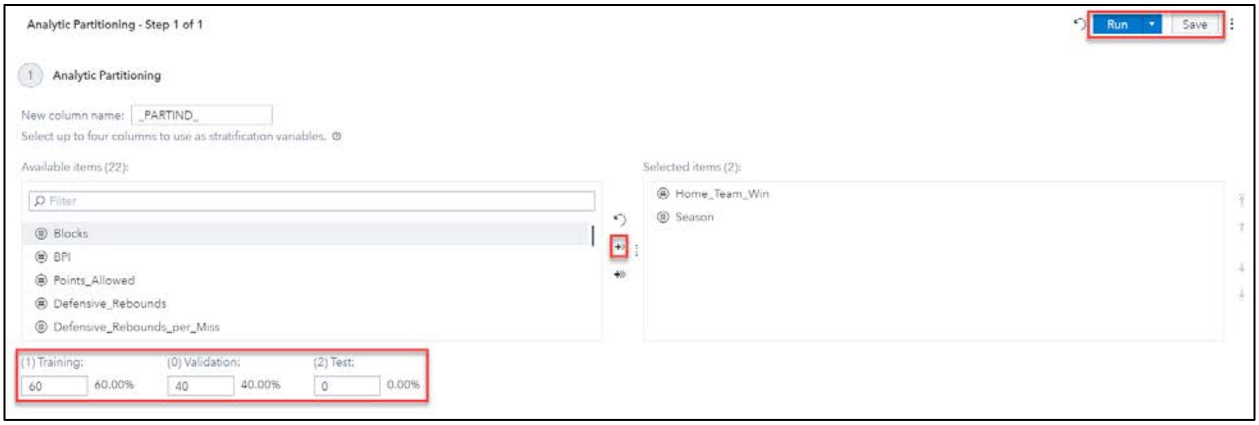
Profiling the data allows the user to quickly identify any unexpected missingness (yes, we just made up a new word), and view distributions, outliers, etc. Once the data is profiled, several data management options are made available in the interface. Options include format columns, transform columns, row transformations, etc. For the purpose of this project, a data partition column must be added to the data. Under the Row Transformations section, click and drag Analytic Partitioning onto the white space. This will bring up several partitioning options that can be seen in the figure below.

PARTITIONING

A quick note to those unfamiliar with predictive modeling and the need for partitioned data: Predictive models can be very powerful and do an exceptional job of finding unobserved relationships within the data – and build those relationships into the data. The problem with this is that the relationships the models find might only exist in the current dataset used to build the model and may not be representative of other data the model will score in the future. This is typically referred to as overfitting a model. To help mitigate this issue it is important to build a model on a portion of the data, and then validate the model's performance against data it has never seen before. For this purpose, the user will almost always want to either partition the data prior to modeling or have a holdout sample of the data that that can be used for model validation.

To partition the data, simply click on the column to include in the partition and then click on the plus sign with one open carrot highlighted above.

One final note about data partitioning: Any time the user takes a sample for model validation, the sample should be representative of what the model is attempting to predict. For the purpose of this exercise, basketball game statistics are being utilized to predict the winner for each game. The target variable to attempt to predict is Home_Team_Win, which is a binary indicator, with 1 if the home team won and 0 if the home team lost. Furthermore, because the data includes multiple seasons, it is a good idea to add this to the partitioning logic as well. This will allow the users to take a stratified random sample on these two variables, which basically means that the unique values of both Season and Home_Team_Win will be equally represented in the data partition.



Analytic Partitioning - Step 1 of 1

1 Analytic Partitioning

New column name: _PARTIND_

Select up to four columns to use as stratification variables.

Available items (22):

- Blocks
- BPI
- Points_Allowed
- Defensive_Rebounds
- Defensive_Rebounds_per_Miss

Selected items (2):

- Home_Team_Win
- Season

(1) Training:		(0) Validation:		(2) Test:	
60	60.00%	40	40.00%	0	0.00%

Figure 5: Data Partition

After the desired variables are selected for partitioning, the user can also set up the partition controls. Because this is a smaller dataset, you will wise a 60 – 40 split between training (the data the model is built on) and validation data. Once the data partition is ready, simply click Run and then Save. To avoid creating a new table after partitioning the data, make sure to click the Replace Table icon underneath the table name.

Once the user has prepped and cleaned the data, the user is ready to explore and visualize the data that will be used. After clicking on the hamburger menu at the top left of the screen, choose "Explore and Visualize".

SAS VISUAL ANALYTICS

When navigating from SAS Data Studio to SAS Visual Analytics, click on the "New Report" option. Now that there is a new report to work with, select the data that was just created in Data Studio.

Click on the data icon on the left vertical ribbon and click Add Data Source. Select the dataset from the available data sources and click "OK".

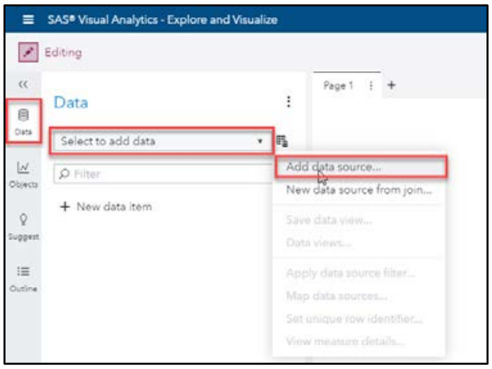
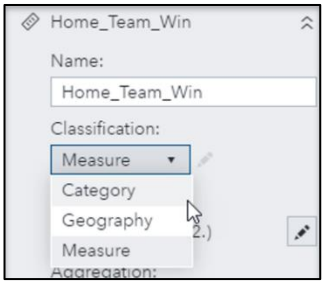


Figure 6: Add Data Source

ADDITIONAL DATA CLEANING

If the user forgot to reformat the target variable from continuous to categorical before moving over to Visual Analytics, the user can still perform several data cleaning capabilities in Visual Analytics.



Home_Team_Win

Name:

Home_Team_Win

Classification:

Measure ▾

Category

Geography

Measure

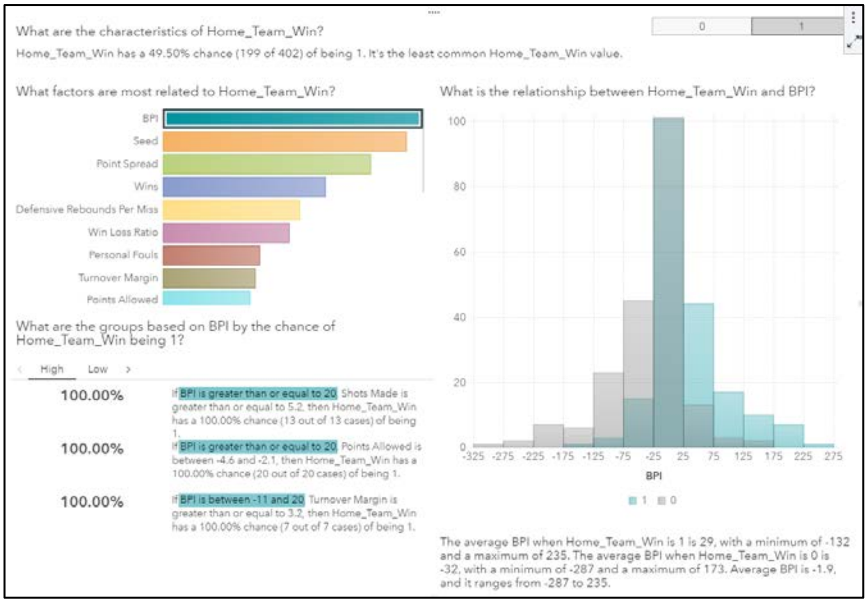
2.)

Aggregation:

Figure 7: Change Category

Scroll down to the column that needs to be reformatted. In this case, Home_Team_Win must be changed from continuous to categorical. Hover the mouse over the variable and click on the drop-down arrow. Next, expand the Classification value and select "Category". The user should notice that Home_Team_Win is updated to classification and is grouped with the partition variable.

Now that the target variable has been converted to categorical, the next step is to explore the relationship the target variable has with the predictor variables. There are two ways to perform this analysis. The user can either right click on the target and select "Explain" or click on the "Objects" icon in the vertical ribbon on the left side of the screen and double-click the Automated Explanation. This analysis, performed either way, will provide a good starting point from a modeling perspective. The software will highlight which variables have the highest predictive power on the target as well as provide some simple business logic that explains how to get the purest splits on the target variable.



What are the characteristics of Home_Team_Win?

Home_Team_Win has a 49.50% chance (199 of 402) of being 1. It's the least common Home_Team_Win value.

What factors are most related to Home_Team_Win?

What are the groups based on BPI by the chance of Home_Team_Win being 1?

High Low

100.00% If BPI is greater than or equal to 20 Shots Made is greater than or equal to 5.2, then Home_Team_Win has a 100.00% chance (13 out of 13 cases) of being 1.

100.00% If BPI is greater than or equal to 20 Points Allowed is between -4.6 and -2.1, then Home_Team_Win has a 100.00% chance (20 out of 20 cases) of being 1.

100.00% If BPI is between -11 and 20 Turnover Margin is greater than or equal to 3.2, then Home_Team_Win has a 100.00% chance (7 out of 7 cases) of being 1.

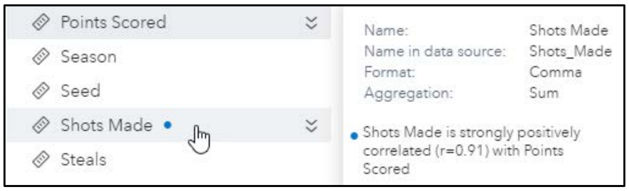
What is the relationship between Home_Team_Win and BPI?

The average BPI when Home_Team_Win is 1 is 29, with a minimum of -132 and a maximum of 235. The average BPI when Home_Team_Win is 0 is -32, with a minimum of -287 and a maximum of 173. Average BPI is -1.9, and it ranges from -287 to 235.

Figure 8: Explain Data

CORRELATIONS

Another important step in building an analytics model is to make sure the model assumptions are met up front. A common assumption for many modeling types is that the predictor variables are not highly correlated with one another. This assumption is known as multicollinearity. Visual Data Mining and Machine Learning allows users to quickly tell which variables may have correlation issues.



Points Scored

Season

Seed

Shots Made ●

Steals

Name:	Shots Made
Name in data source:	Shots_Made
Format:	Comma
Aggregation:	Sum

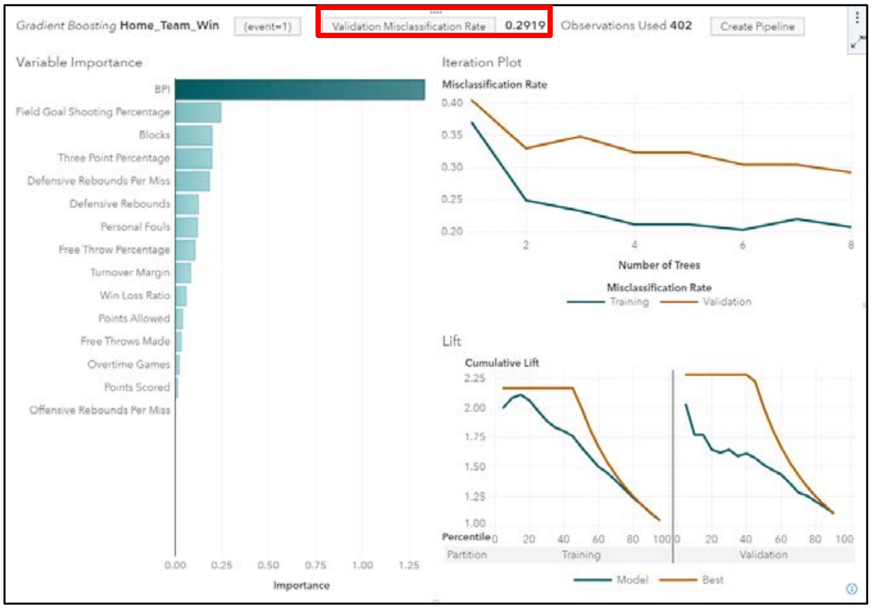
● Shots Made is strongly positively correlated (r=0.91) with Points Scored

Figure 9: Variable Correlations

After double-clicking on a variable listed under the "Data" icon on the left side of the screen, Visual Data Mining and Machine Learning will automatically run correlation tests on the highlighted column against all the other variables in the data. Highly correlated variables will be flagged with a blue dot. As seen in the figure, by simply clicking on the variable Points Scored, one can easily see that Shots Made is highly correlated with Points Scored. By taking this small extra step, one can quickly assess some important model assumptions that should never be overlooked.

BASELINE MODEL

Now that the data have been explored and some preliminary analysis has been run, a baseline model can be built. Visual Data Mining and Machine Learning provides several predictive modeling options. For this exercise, you will start with a Gradient Boosting model. Under the Objects tab on the left side of the screen, scroll down to SAS Visual Data Mining and Machine Learning and double-click on the "Gradient Boosting" node. Once the default template is loaded into the workspace click on the "Roles" tab on the right of the screen and select the Response, Predictors, and Partition ID variables. The moment the variables are selected, the Gradient Boosting model will automatically run and produce the model and model assessment charts and values. Each model has several control mechanisms available so that the model can be custom-built to the desired specifications. The model built here uses the standard autotuning feature provided by Visual Data Mining and Machine Learning.



Gradient Boosting Home_Team_Win (event=1) Validation Misclassification Rate 0.2919 Observations Used 402 Create Pipeline

Figure 10: Model Output

Because you are predicting a binary outcome with a nearly even distribution of the target, the misclassification rate of the validation dataset was selected as the model selection criteria. SAS provides several model selection criteria that can be selected by clicking the box at the top highlighted in red in the figure above.

If the user finds a model that they desire to use as a champion model, the model can be registered directly from SAS Visual Analytics. Simply click the ellipsis at the top right corner in the image above and click Register Model. This will save the model and make it available for others to work with.

However, most of the time the user will want to build a base-level model and then see if they can improve upon the model performance. To do this click on the ellipsis in the top right corner and select "Create Pipeline". This will save the model and transfer it to SAS Model Studio.

SAS MODEL STUDIO

SAS Model Studio is where a lot of the meat exists within Visual Data Modeler and Machine Learning. This environment is like the SAS Viya version of SAS Enterprise Miner. Multiple supervised and unsupervised models can be built within one pipeline, model performances can be compared in one run, and even ensemble models can be created.

Similar to Visual Analytics, Model Studio allows the user to custom build each model available. To add a new model to the pipeline click on the ellipsis in the "Data Preparation" node in the pipeline and click Add Child Node -> Supervised Learning and then select the desired model. Each model built will be added to the pipeline and made available in the final model comparison.

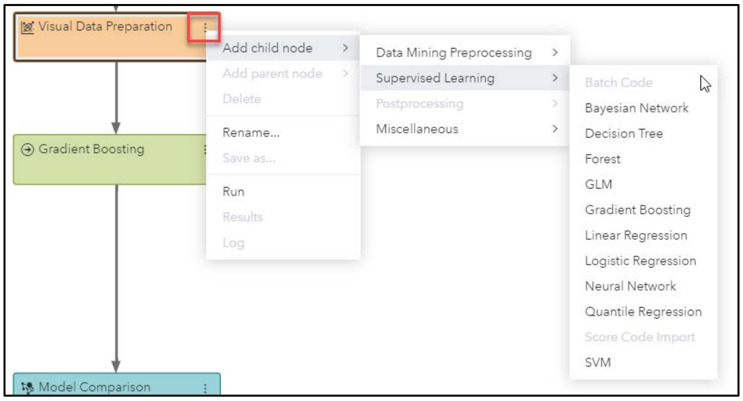


Figure 11: Adding Supervised Learning Models

To add an ensemble model to the pipeline, simply open the menu in any of the model nodes and select Add Child Node -> Postprocessing -> Ensemble. The user can then add any combination of the other models to include as part of the ensemble.

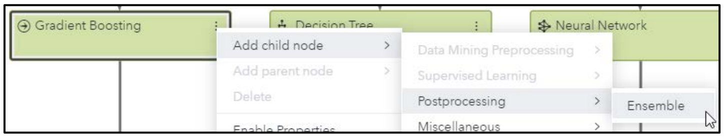
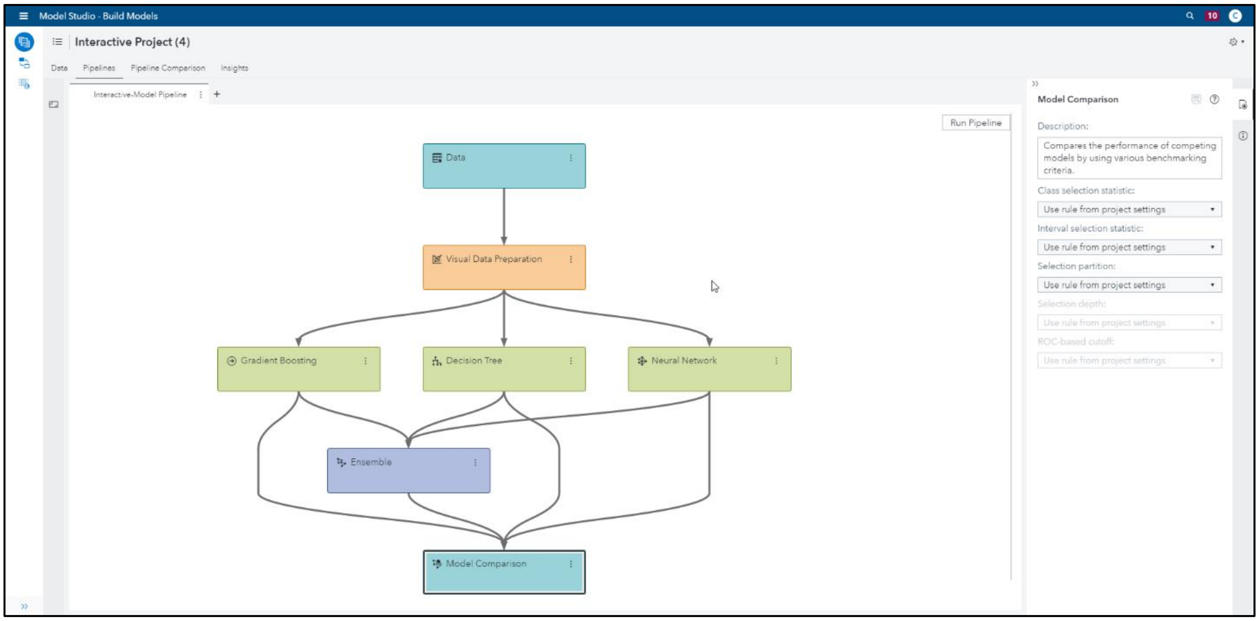


Figure 12: Ensemble Model

After the models to run against the validation data are set up, select the model selection criteria. Select the "Model Comparison" node and expand the Model Comparison Options on the right side of the workspace. Here several model selection criteria will be available. Again, the Validation Misclassification Rate is being used. Once the pipeline is set the user can simply click on the "Run Pipeline" button in the top right corner of the workspace. The entire pipeline will run.



Model Comparison

Description:

Compares the performance of competing models by using various benchmarking criteria.

Class selection statistic: Use rule from project settings

Interval selection statistic: Use rule from project settings

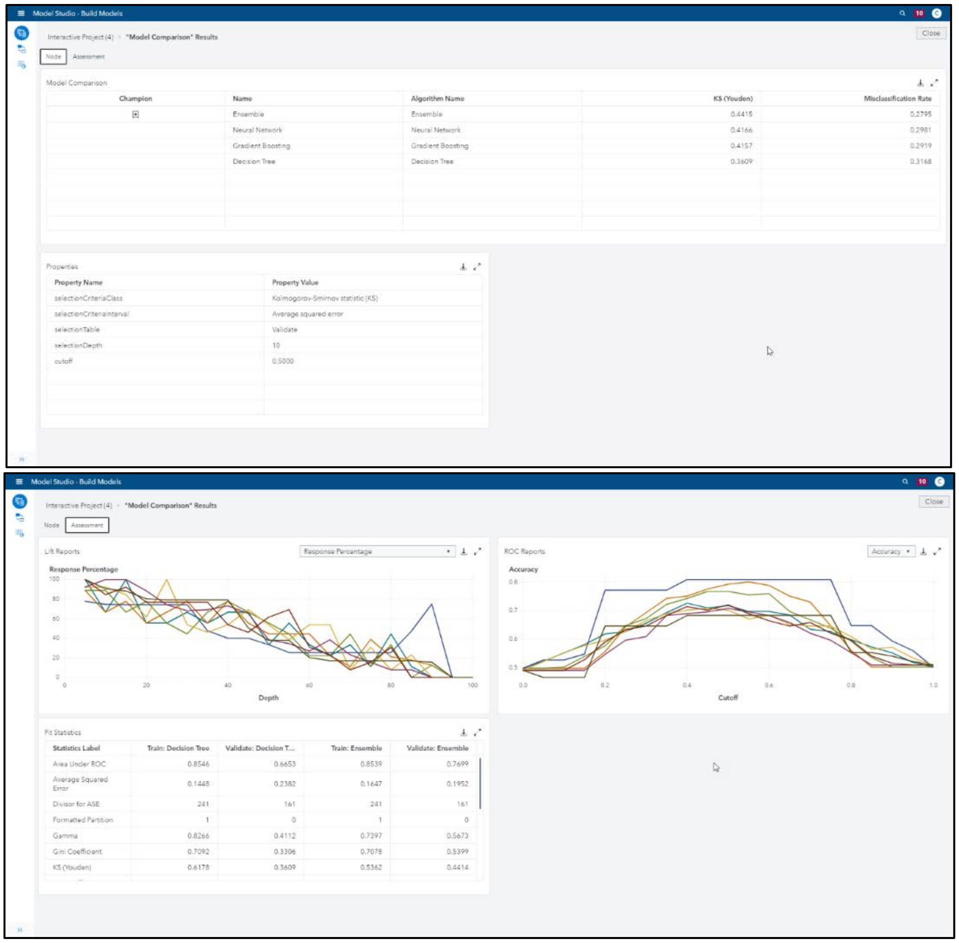
Selection partition: Use rule from project settings

Selection depth: Use rule from project settings

ROC-based cutoff: Use rule from project settings

Figure 13: VDMML Pipeline

When the pipeline successfully executes, click on the "Model Comparison" ellipsis and select "Results". This will open the complete results from the run with the champion model highlighted and all the fit statistic plots. As seen in the figure below, the Ensemble model had the lowest misclassification rate.



Interactive Project (4) › "Model Comparison" Results

Model Comparison

Champion	Name	Algorithm Name	KS (Youden)	Misclassification Rate
☒	Ensemble	Ensemble	0.4415	0.2795
	Neural Network	Neural Network	0.4166	0.2981
	Gradient Boosting	Gradient Boosting	0.4157	0.2919
	Decision Tree	Decision Tree	0.3609	0.3168

Properties

Property Name	Property Value
selectionCriteriaClass	Kolmogorov-Smirnov statistic (KS)
selectionCriteriaInterval	Average squared error
selectionTable	Validate
selectionDepth	10
cutoff	0.5000

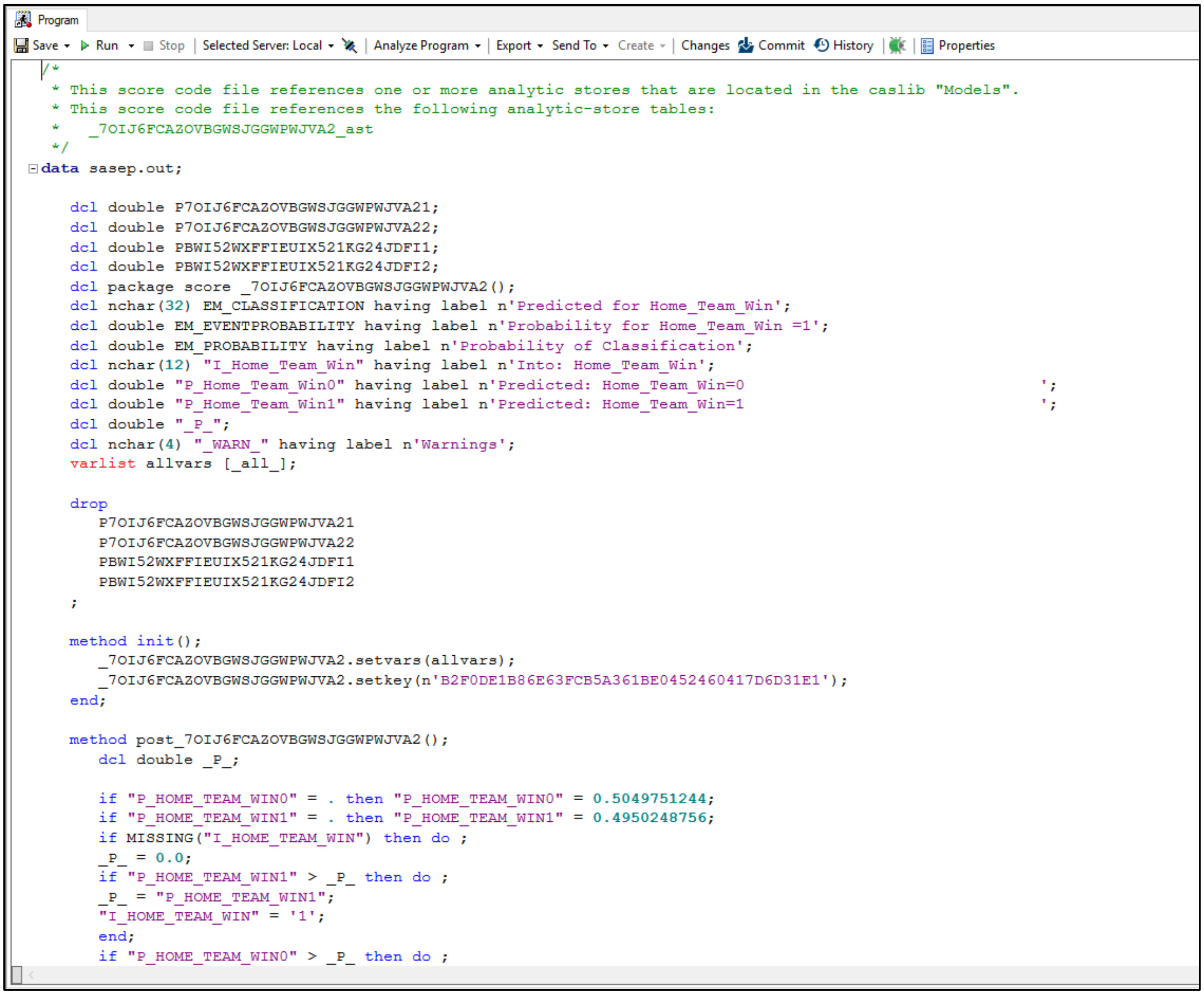
Interactive Project (4) › "Model Comparison" Results

Fit Statistics

Statistics Label	Train: Decision Tree	Validate: Decision T...	Train: Ensemble	Validate: Ensemble
Area Under ROC	0.8546	0.6653	0.8539	0.7699
Average Squared Error	0.1448	0.2382	0.1647	0.1952
Divisor for ASE	241	161	241	161
Formatted Partition	1	0	1	0
Gamma	0.8266	0.4112	0.7397	0.5673
Gini Coefficient	0.7092	0.3306	0.7078	0.5399
KS (Youden)	0.6178	0.3609	0.5362	0.4414

Figure 14: Pipeline Results

Once a champion model is selected, the user can either register the model as previously discussed or save the actual score code for the model. Throughout this entire process not a single line of code was typed to build the final ensemble model. SAS Visual Data Mining and Machine Learning will create the entire code for the user. Simply click the ellipsis of the node for the champion model and select 〝Download Score Code〞. Visual Data Mining and Machine Learning generates the SAS code without the user typing a single line of code.



```
/*
 * This score code file references one or more analytic stores that are located in the caslib "Models".
 * This score code file references the following analytic-store tables:
 *    _7OIJ6FCAZOVBGWSJGGWPWJVA2_ast
 */
data sasep.out;

   dcl double P7OIJ6FCAZOVBGWSJGGWPWJVA21;
   dcl double P7OIJ6FCAZOVBGWSJGGWPWJVA22;
   dcl double PBWI52WXFFIEUIX521KG24JDFI1;
   dcl double PBWI52WXFFIEUIX521KG24JDFI2;
   dcl package score _7OIJ6FCAZOVBGWSJGGWPWJVA2();
   dcl nchar(32) EM_CLASSIFICATION having label n'Predicted for Home_Team_Win';
   dcl double EM_EVENTPROBABILITY having label n'Probability for Home_Team_Win =1';
   dcl double EM_PROBABILITY having label n'Probability of Classification';
   dcl nchar(12) "I_Home_Team_Win" having label n'Into: Home_Team_Win';
   dcl double "P_Home_Team_Win0" having label n'Predicted: Home_Team_Win=0                              ';
   dcl double "P_Home_Team_Win1" having label n'Predicted: Home_Team_Win=1                              ';
   dcl double "_P_";
   dcl nchar(4) "_WARN_" having label n'Warnings';
   varlist allvars [_all_];

   drop
      P7OIJ6FCAZOVBGWSJGGWPWJVA21
      P7OIJ6FCAZOVBGWSJGGWPWJVA22
      PBWI52WXFFIEUIX521KG24JDFI1
      PBWI52WXFFIEUIX521KG24JDFI2
   ;

   method init();
      _7OIJ6FCAZOVBGWSJGGWPWJVA2.setvars(allvars);
      _7OIJ6FCAZOVBGWSJGGWPWJVA2.setkey(n'B2F0DE1B86E63FCB5A361BE0452460417D6D31E1');
   end;

   method post_7OIJ6FCAZOVBGWSJGGWPWJVA2();
      dcl double _P_;

      if "P_HOME_TEAM_WIN0" = . then "P_HOME_TEAM_WIN0" = 0.5049751244;
      if "P_HOME_TEAM_WIN1" = . then "P_HOME_TEAM_WIN1" = 0.4950248756;
      if MISSING("I_HOME_TEAM_WIN") then do ;
      _P_ = 0.0;
      if "P_HOME_TEAM_WIN1" > _P_ then do ;
      _P_ = "P_HOME_TEAM_WIN1";
      "I_HOME_TEAM_WIN" = '1';
      end;
      if "P_HOME_TEAM_WIN0" > _P_ then do ;
```

Figure 15: Export Model Score Code

CONCLUSION

SAS Visual Data Mining and Machine Learning is a true end-to-end solution for both technical and non-technical stake holders. The flexible user interface allows for rapid data exploration and modeling. What used to take weeks to accomplish can now be done with just a few clicks of the mouse.

For further insights about the power of SAS Visual Data Mining and Machine Learning, please see the additional resources provided below.

RECOMMENDED READING

- https://www.zencos.com/capabilities/advanced-analytics/
- https://www.sas.com/en_us/software/visual-data-mining-machine-learning.html

CONTACT INFORMATION

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

Chris St. Jeor
Zencos
cst.jeor@zencos.com
www.zencos.com