Getting Started with SAS® Viya™ Data Mining and Machine Learning

Early Adopter Software

THIS DOCUMENTATION FOR AN EARLY ADOPTER PRODUCT IS A PRELIMINARY DRAFT AND IS PROVIDED BY SAS INSTITUTE INC. (“SAS”) ON AN "AS IS" BASIS WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, IMPLIED WARRANTIES OF MERCHANTABILITY AND/OR FITNESS FOR A PARTICULAR PURPOSE. SAS does not warrant that this documentation is complete, accurate, or similar to that which may be released to the general public, or that any such documentation will be released. The company shall not be liable whatsoever for any damages arising out of the use of this documentation, including any direct, indirect, or consequential damages. SAS reserves the right to alter or abandon use of this documentation at any time.

Audience

Data scientists, experienced statisticians, data miners, engineers, researchers, and scientists, who need to analyze large and complex data and build predictive models using a modern, powerful, and customizable in-memory programming language.

What is Machine Learning?

Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look.
Benefits of Using Machine Learning

- You can reduce the time it takes to transform raw data into decisions through an integrated and complete machine learning environment that integrates all necessary components of machine learning pipelines.
- Data scientists can run more model-building scenarios in the same environment, which improves their productivity.
- Data scientists can build sophisticated machine learning pipelines with far less code because the integrated modern machine learning capabilities are available in a higher-level programming language.
- Data science teams can collaborate more easily through a web-based interface and a common machine learning platform.
- Organizations can extend the talent pool of data scientists by leveraging existing SAS skills for machine learning projects.

Workflow for Machine Learning

1. **Identify** your data sets, load the data sets into SAS Cloud Analytic Services, and perform any data manipulation.
2. **Explore** the data by creating graphs that visualize any anticipated relationships, unanticipated trends, and anomalies in order to gain understanding and ideas.
3. **Prepare** the data for model building by splitting the data into training, validation, and test data.
4. **Model** the data by using the analytical techniques to search for a combination of the data that reliably predicts a desired outcome.
5. **Assess** the data by evaluating the usefulness and reliability of the findings from the data mining process.
6. **Score** new data. Assess and monitor the effectiveness of the models.

Access Data, Create Formats, Create Input Data Set

**Where Can I Get the Data?**

The case study begins with two SAS data sets that are available from the following location in GitHub:

https://github.com/sassoftware/sas-viya-programming/tree/master/chicago

The two data sets are named census.sas7bdat and crimes.sas7bdat. Later in the case study, we will show how to join the two data sets to integrate the data as follows:

- The crimes data set identifies the offenses and includes a community area value that identifies the vicinity of the offense.
The community area value maps to the census data set that contains six socioeconomic indicators of public health, a “hardship index,” and per capita income.

Data Provenance

City of Chicago Crime Data (2001 to present)
The data set that is used in the examples is a subset of the original data set that is available from the City of Chicago. The date range for the sample is January 1, 2001, to February 15, 2016.

The original data is available from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2.

City of Chicago Census Data (2008 to 2012)
The data set is available from https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2.

Disclaimer: SAS may reference other websites or content or resources for use at Customer’s sole discretion. SAS has no control over any websites or resources that are provided by companies or persons other than SAS. Customer acknowledges and agrees that SAS is not responsible for the availability or use of any such external sites or resources, and does not endorse any advertising, products, or other materials on or available from such websites or resources. Customer acknowledges and agrees that SAS is not liable for any loss or damage that may be incurred by Customer or its end users as a result of the availability or use of those external sites or resources, or as a result of any reliance placed by Customer or its end users on the completeness, accuracy, or existence of any advertising, products, or other materials on, or available from, such websites or resources.

Very First Steps

1 If you have not already, download the two data sets. Put them in a directory that is accessible to the SAS Workspace Server.

2 Sign in to SAS Studio.

   a Use the Server Files and Folders section to navigate to the directory that has the two data sets.
   b Right-click on the directory and select Create ➔ Library. Specify chicago as the name.

3 Use the Libraries section to confirm that the new library is listed.

4 Check that you have a CAS session.

   a Enter the following line in the code editor and click to run the code.
Click the **Code** tab. A note like the following indicates that you have a CAS session.

NOTE: Session CASAUTO is ACTIVE using port 5570 and host cloud.example.com for user sasdemo. The session UUID is 2861d3a4-779b-674c-bb46-31f46a1dafa0.

**Create a Format**

One of the frequently used variables in the crimes data set is the Fbi_code column. The codes in this column correspond to the uniform crime reporting codes as described at http://gis.chicagopolice.org/clearmap_crime_sums/crime_types.html.

Instead of using the codes such as "06" in displays to indicate larceny, we will create a SAS format that maps the code to the more descriptive text, Larceny. Enter the following code in the **Code** tab and then click **Run**.

**Note:** The CASFMTLIB= option is introduced in the SAS Viya 3.0 release. It is required to add the format to SAS Cloud Analytic Services.

```sas
proc format casfmtlib="chicagofmts" sessref=casauto;
  value $fbi
    '01A' = 'Homicide 1st & 2nd Degree'
    '02' = 'Criminal Sexual Assault'
    '03' = 'Robbery'
    '04A' = 'Aggravated Assault'
    '04B' = 'Aggravated Battery'
    '05' = 'Burglary'
    '06' = 'Larceny'
    '07' = 'Motor Vehicle Theft'
    '08A' = 'Simple Assault'
    '08B' = 'Simple Battery'
    '09' = 'Arson'
    '10' = 'Forgery & Counterfeiting'
    '11' = 'Fraud'
    '12' = 'Embezzlement'
    '13' = 'Stolen Property'
    '14' = 'Vandalism'
    '15' = 'Weapons Violation'
    '16' = 'Prostitution'
    '17' = 'Criminal Sexual Abuse'
    '18' = 'Drug Abuse'
    '19' = 'Gambling'
    '20' = 'Offenses Against Family'
    '22' = 'Liquor License'
    '24' = 'Disorderly Conduct'
    '26' = 'Misc Non-Index Offense'
  ;
  value suntosat
    1 = 'Sunday'
    2 = 'Monday'
    3 = 'Tuesday'
    4 = 'Wednesday'
    5 = 'Thursday'
    6 = 'Friday'
    7 = 'Saturday';
run;
```
Load the Data Into SAS Cloud Analytic Services

Enter the following on the Code tab and then click ✈️.

```sas
proc casutil;
  load data=chicago.census;
  load data=chicago.crime;
  contents casdata="crime";
  contents casdata="census";
quit;
```

If you did not assign the name `chicago` when you created the SAS library, substitute the value that you used.

Merge the Data Sets

To get a more complete picture of each crime, the census data is joined to the crime table.

```sas
libname mycas cas sessref=casauto;

data mycas.crimeCensus/ sessref=casauto;
  merge mycas.crime(in=in1) mycas.census(in=in2);
  by community_area;
  if in1;
run;

proc casutil;
  contents casdata="crimeCensus";
quit;
```
Table and Column Information for the crimeCensus Table

The CASUTIL Procedure

Table Information for Caslib CASUSER(<CasLib>)

<table>
<thead>
<tr>
<th>Table Name</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
<th>NLS encoding</th>
<th>Created</th>
<th>Last Modified</th>
<th>Promoted Table</th>
<th>Repeated Table</th>
<th>View</th>
<th>Compressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRIMECENSUS</td>
<td>105951</td>
<td>31</td>
<td>utf-8</td>
<td>30 Jun 2016 13:04:41</td>
<td>30 Jun 2016 13:04:41</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Detail Information for crimeCensus in Caslib CASUSER(<CasLib>).

<table>
<thead>
<tr>
<th>Node</th>
<th>Number of Blocks</th>
<th>Active Blocks</th>
<th>Blocks</th>
<th>Memory Mapped</th>
<th>Blocks Unmapped</th>
<th>Memory Unmapped</th>
<th>Blocks Allocated</th>
<th>Memory Allocated</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>192</td>
<td>192</td>
<td>105951</td>
<td>0</td>
<td>0</td>
<td>127</td>
<td>33913762</td>
<td>65</td>
</tr>
</tbody>
</table>

Column Information for CRIMECENSUS in Caslib CASUSER(<CasLib>),

<table>
<thead>
<tr>
<th>Column</th>
<th>Type</th>
<th>Length</th>
<th>Formatted Length</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>date</td>
<td>double</td>
<td>B</td>
<td>10</td>
<td>MMDXYY</td>
</tr>
<tr>
<td>fbi_code</td>
<td>char</td>
<td>25</td>
<td>25</td>
<td>$FBI</td>
</tr>
<tr>
<td>arrest_code</td>
<td>B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>percent_over_65</td>
<td>double</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>percent_of_housing_crowded</td>
<td>double</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>percent_aged_65_without_high_sch</td>
<td>double</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>percent_aged_16_unemployed</td>
<td>double</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>percent_households_below_poverty</td>
<td>double</td>
<td>12</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Explore Data Using ODS Graphics

Before performing data analysis, it is important to explore and survey the data. With larger data sets, that is done by summarizing the data and then either printing the summary table or creating a plot.

Plot the Overall Arrest Rate

/* overall arrest rate */
ods graphics / width=5in antialiasmax=5600;  
proc sgplot data=mycas.crimeCensus;  
title "Overall Arrest Rate";  
vbar arrest;  
run;  
run;
Plot the Crime and Arrest Rate by Offense

This plot helps visualize that some types of crime, such as Larceny and Simple Battery, occur much more often than other types of crime.

```sas
ods graphics / width=8in antialiasmax=5600;
proc sgplot data=mycas.crimeCensus;
    vbar fbi_code / categoryorder=respdesc group=arrest;
    xaxis display=(nolabel);
run;
```
To see the arrest rate as a percentage of crimes for the top 10 crimes, you can run the following code:

```sas
proc mdsummary data=mycas.crimeCensus;
  var arrest_code;
  groupby fbi_code;
  output out=mycas.crimeGrouped;
run;
```

Because the MDSUMMARY procedure processes numeric variables only, the arrest_code variable that is coded as numeric 0 and 1 is used. The rows are displayed below:
## Top Ten Crimes and Arrest Rate as a Percentage

<table>
<thead>
<tr>
<th>FBI Code</th>
<th>Crimes</th>
<th>Arrests</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larceny</td>
<td>21433</td>
<td>2792</td>
<td>13.03</td>
</tr>
<tr>
<td>Simple Battery</td>
<td>16593</td>
<td>4083</td>
<td>24.61</td>
</tr>
<tr>
<td>Vandalism</td>
<td>12049</td>
<td>894</td>
<td>7.42</td>
</tr>
<tr>
<td>Drug Abuse</td>
<td>11440</td>
<td>11369</td>
<td>99.38</td>
</tr>
<tr>
<td>Misc Non-Index Offense</td>
<td>11105</td>
<td>4508</td>
<td>40.59</td>
</tr>
<tr>
<td>Burglary</td>
<td>6163</td>
<td>411</td>
<td>6.67</td>
</tr>
<tr>
<td>Motor Vehicle Theft</td>
<td>5001</td>
<td>486</td>
<td>9.72</td>
</tr>
<tr>
<td>Simple Assault</td>
<td>4897</td>
<td>1049</td>
<td>21.42</td>
</tr>
<tr>
<td>Robbery</td>
<td>3991</td>
<td>418</td>
<td>10.47</td>
</tr>
<tr>
<td>Fraud</td>
<td>3029</td>
<td>654</td>
<td>21.59</td>
</tr>
</tbody>
</table>

## Plot Larceny and Battery by Hour

The two largest categories are Larceny and Simple Battery. Is there any relationship between time of day and day of the week and having the crime result in an arrest?

```plaintext
data mycas.larBatByHr / sessref=casauto;
  set mycas.crimeCensus;
  format dow suntosat.;
  if fbi_code in ('06', '08B');
  dow=weekday(date); /* returns 1=Sunday... */
  h=hour(timestamp);
run;

proc sort data=mycas.larBatByHr out=work.sorted;
  by dow;
run;

proc sgplot data=work.sorted;
  heatmap x=dow y=h / colorresponse=arrest_code discretex;
  yaxis label="Hour of Day";
  xaxis discreteorder=data display=(nolabel);
run;
```
Heat Map of Crime by Hour

Plot Crime by per Capita Income

The documentation for the census data set indicates the following national averages for a number of statistics. According to United States Census Bureau 2008-2012 American Community Survey 5-year estimates:

- 3.2% of occupied housing units in the US had more than one person per room
- 10.9% of households were living below the federal poverty level
- 9.3% of persons aged 16 years or older in the labor force were unemployed
- 14.2% of persons aged 25 years or older did not have a high school diploma
- 37.2% of the population was under 18 or more than 64 years of age
- per capita income was $28,051

```
proc sgplot data=mycas.crimeCensus;
    title 'Crime incidence by Per Capita Income';
    histogram per_capita_income;
    reline 28051 / axis=x label='U.S. Average';
    xaxis label='Per Capita Income';
run;
```

The plot shows a cluster of reported crime at and below the national average.
Plot Crime and Arrest over Time

By plotting a time series of the number of reported crimes and arrests, we can see an annual trend and a seasonal trend. We can also see how the number of reported crimes is trending down in recent years compared to the early 2000s.

```plaintext
proc mdsummary data=mycas.crimeCensus;
   groupby date;
   var arrest_code;
   output out=mycas.ts;
run;

proc sgplot data=mycas.ts;
   title "Times Series of Crime and Arrest";
   series x=date y=_sum_ / legendlabel='Arrests' connectorder=xaxis;
   series x=date y=_nobs_ / legendlabel='Reported Crimes' connectorder=xaxis;
run;
```
Plot Crime Incidents by Hardship Index

The kernel density estimate shows an increase in reported crime in areas with a hardship index that exceeds 50.

```
proc sgplot data=mycas.crimeCensus;
   title 'Percentage of Crime by Hardship Index';
   histogram hardship_index;
   density hardship_index / type=kernel;
run;
```
Perform Data Prep

Split the Data into Training, Validation, and Test

The **PARTITION** procedure is used to perform stratified sampling.

```plaintext
proc partition data=mycas.crimeCensus partind seed=9878 sampct=30 sampct2=10;
    target arrest;
    output out=mycas.cwpart copyvars=_all_;
run;
```

- This **PARTIND** option adds a partition indicator column to the output table. The column has values of 0, 1, or 2 that correspond to the training, validation, and testing partitions.
- The **SEED=** option specifies 9878 as the random seed to be used in the partitioning process.
- The **TARGET** statement causes the procedure to perform stratified sampling and maintain the ratio of arrest in each partition.
- The sampling is controlled as follows (the value for the partition indicator column is shown in parentheses):
  - **SAMPPCT=30** requests that 30% of the input data be included in the validation partition (1).
  - **SAMPPCT2=10** requests that 10% of the input data be included in the testing partition (2).
  - the balance, 60%, is used for the training partition (0).
- The **OUTPUT** statement requests that the sampled data be stored in a table named mycas.cwpart and the **COPYVARS= _ALL_** option specifies that all variables are copied from mycas.crimeCensus to mycas.cwpart.
The table mycas.cwpart is used as the input table for creating the models in the section Create the Models on page 15.

**Stratified Sampling Results**

The PARTITION Procedure

<table>
<thead>
<tr>
<th>Index</th>
<th>Arrest</th>
<th>Number of Obs</th>
<th>Sample Size 1</th>
<th>Sample Size 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>false</td>
<td>74283</td>
<td>22285</td>
<td>7428</td>
</tr>
<tr>
<td>1</td>
<td>true</td>
<td>31668</td>
<td>9500</td>
<td>3167</td>
</tr>
</tbody>
</table>

Output CAS Tables

<table>
<thead>
<tr>
<th>CAS Library</th>
<th>Name</th>
<th>Number of Rows</th>
<th>Number of Columns</th>
</tr>
</thead>
<tbody>
<tr>
<td>CASUSERHDFS</td>
<td>CWPART</td>
<td>10595</td>
<td>32</td>
</tr>
</tbody>
</table>

**Check the Proportion of the Sampling**

The MDSUMMARY procedure creates basic descriptive statistics on data in the server. We want to verify that the arrest variable was stratified across the three partitions.

```sas
proc mdsummary data=mycas.cwpart;
    groupby _partind_;
    var arrest_code;
    output out=mycas.split;
run;
proc print data=mycas.split;
run;
```

By looking at the _Mean_ column we can verify that all three rows have a value of about 0.29. We can also verify the percentage of observations in each partition by looking at the _NObs_ column.

**Partial Results for the Mycas.split Table**

<table>
<thead>
<tr>
<th>Obs</th>
<th>Partind</th>
<th>Partind_f</th>
<th>Column</th>
<th><em>Min</em></th>
<th><em>Max</em></th>
<th><em>NOb</em></th>
<th><em>NMiss</em></th>
<th><em>Mean</em></th>
<th><em>Sum</em></th>
<th><em>Std</em></th>
<th><em>StdErr</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>arrest_code</td>
<td>0</td>
<td>1</td>
<td>10595</td>
<td>0</td>
<td>0.2998945624</td>
<td>3167</td>
<td>0.0044476339</td>
<td>f</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>arrest_code</td>
<td>0</td>
<td>1</td>
<td>63571</td>
<td>0</td>
<td>0.2988941948</td>
<td>19001</td>
<td>0.0018156177</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>arrest_code</td>
<td>0</td>
<td>1</td>
<td>31785</td>
<td>0</td>
<td>0.298883121</td>
<td>9500</td>
<td>0.0025676843</td>
<td>f</td>
</tr>
</tbody>
</table>

**Explore the Cardinality**

The CARDINALITY procedure determines a variable’s cardinality (the number of levels the variable has) or limited cardinality in SAS Viya 3.0. You can view the results of the CARDINALITY procedure to decide which variables to use as classification or interval variables for model building.

```sas
proc cardinality data=mycas.cwpart outcard=mycas.card maxlevels=20;
    var _numeric_;
    var _char_;
run;
proc print data=mycas.card;
run;
```
Create the Models

Now you can build candidate models based on the training and validation partition.

Create Macro Variables

It is helpful to set up macro variables for repetitive code. In this example, we create macros for the commonly used user-supplied values.

```sas
%let dset=mycas.CWpart;
%let outdir=--;
%let target=arrest_code;
%let nom_input=fbi_code location_description domestic beat district ward community_area;
%let int_input= percent: per_capita_income hardship_index;
```

- The DSET macro variable represents the input table, mycas.CWpart. Input data must be an in-memory table that is accessible in your CAS session. You must refer to this table by using a two-level name. The first level must be a CAS engine libref and the second level must be the table name. In this example, `mycas.` is the CAS engine libref, so we begin the input table with `mycas.`.
- The OUTDIR macro variable is used to specify an operating system directory for the SAS scoring code that analytic procedures create. The directory must be accessible to the SAS Workspace Server.
- The TARGET macro variable represents the variable whose values the procedures try to predict. Arrest_code was originally the character variable Arrest in the input data. We transformed it to a numeric variable.
- The NOM_INPUT and INT_INPUT macro variables are used to specify the columns to treat as nominal and interval values, respectively. Using a macro variable can speed your development time because the columns are specified in one place and referenced multiple times.

Create a Predictive Model

The FOREST procedure creates a predictive model called a forest in SAS Viya. The purpose of a predictive model is to predict a target value from inputs. The following statements train a forest model and score the training data table.

```sas
proc forest data=&dset. ntrees=10 minleafsize=5 outmodel=mycas.model_forest;
    target &target. / level=nominal;
    input &nom_input. / level=nominal;
    input &int_input. / level=interval;
    partition rolevar=_partind_(train='0' validate='1');
    output out=mycas.ap_scored_forest copyvars=(_partind_ &target);
    title "Random Forest";
run;
```
The TARGET, INPUT, PARTITION, and OUTPUT statements perform the same basic function in all of the procedures in this section.

- The TARGET statement names the variable whose values PROC FOREST tries to predict.
- The INPUT statements specify the nominal and interval input variables. We determined the interval and nominal variables from the PROC CARDINALITY results.
- The PARTITION statement assigns roles to the observations in the input table based on the value of the variable _partind_ in that table. Observations whose value is 0 are assigned for training, and observations with a value of 1 are used for validation.
- The OUTPUT statement creates a new data table mycas.ap_scored_forest that is the result of prediction from using the input data and the model. PROC ASSESS uses this table as the input table in Assess the Models.

FOREST Procedure Results

<table>
<thead>
<tr>
<th>The FOREST Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Information</td>
</tr>
<tr>
<td>Number of Trees</td>
</tr>
<tr>
<td>Number of Variables Per Split</td>
</tr>
<tr>
<td>Seed</td>
</tr>
<tr>
<td>Bootstrap Percentage</td>
</tr>
<tr>
<td>Number of Bins</td>
</tr>
<tr>
<td>Number of Input Variables</td>
</tr>
<tr>
<td>Maximum Number of Tree Nodes</td>
</tr>
<tr>
<td>Minimum Number of Tree Nodes</td>
</tr>
<tr>
<td>Maximum Number of Branches</td>
</tr>
<tr>
<td>Minimum Number of Branches</td>
</tr>
<tr>
<td>Maximum Depth</td>
</tr>
<tr>
<td>Minimum Depth</td>
</tr>
<tr>
<td>Maximum Number of Leaves</td>
</tr>
<tr>
<td>Minimum Number of Leaves</td>
</tr>
<tr>
<td>Maximum Leaf Size</td>
</tr>
<tr>
<td>Minimum Leaf Size</td>
</tr>
<tr>
<td>OOB Misclassification Rate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of Observations Read</th>
<th>Training</th>
<th>Validation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Used</td>
<td>63571</td>
<td>31785</td>
<td>95356</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable Importance</th>
<th>importance</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>tbl_code</td>
<td>6451.33</td>
<td>265.54</td>
</tr>
<tr>
<td>Location_Description</td>
<td>33.45</td>
<td>17.3240</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fit Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Trees</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
</tbody>
</table>
Create Decision Trees with Gradient Boosting

The **GRADBOOST** procedure creates a series of decision trees that together form a single predictive model. The output table mycas.ap_scored_gradboost is used by PROC ASSESS in the section **Assess the Models**.

```plaintext
proc gradboost data=&dset. maxdepth=8 minleafsize=5 seed=9878 outmodel=mycas.model_gradboost;
   target &target. / level=nominal;
   input &nom_input. / level=nominal;
   input &int_input. / level=interval;
   partition rolevar=_partind_(train='0' validate='1');
   output out=mycas.ap_scored_gradboost copyvars=(_partind_ &target.);
   title "Gradient Boost";
run;
```
GRADBOOST Procedure Results

Gradient Boost

The GRADBOOST Procedure

Model Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Trees</td>
<td>100</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>0.1</td>
</tr>
<tr>
<td>Subsampling Rate</td>
<td>0.5</td>
</tr>
<tr>
<td>Number of Variables Per Split</td>
<td>14</td>
</tr>
<tr>
<td>Number of Bins</td>
<td>20</td>
</tr>
<tr>
<td>Number of Input Variables</td>
<td>14</td>
</tr>
<tr>
<td>Maximum Number of Tree Nodes</td>
<td>489</td>
</tr>
<tr>
<td>Minimum Number of Tree Nodes</td>
<td>283</td>
</tr>
<tr>
<td>Maximum Number of Branches</td>
<td>2</td>
</tr>
<tr>
<td>Minimum Number of Branches</td>
<td>2</td>
</tr>
<tr>
<td>Maximum Depth</td>
<td>8</td>
</tr>
<tr>
<td>Minimum Depth</td>
<td>8</td>
</tr>
<tr>
<td>Maximum Number of Leaves</td>
<td>245</td>
</tr>
<tr>
<td>Minimum Number of Leaves</td>
<td>142</td>
</tr>
<tr>
<td>Maximum Leaf Size</td>
<td>6129</td>
</tr>
<tr>
<td>Minimum Leaf Size</td>
<td>5</td>
</tr>
<tr>
<td>Seed</td>
<td>9878</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Validation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
<td>63571</td>
<td>31785</td>
<td>95356</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>63571</td>
<td>31785</td>
<td>95356</td>
</tr>
</tbody>
</table>

Variable Importance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Importance</th>
<th>Std Dev Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>fbi_code</td>
<td>447.31</td>
<td>146.72</td>
</tr>
<tr>
<td>Beat</td>
<td>319.47</td>
<td></td>
</tr>
<tr>
<td>Location_Deah</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community_Area</td>
<td>68.2076</td>
<td>2.4899</td>
</tr>
<tr>
<td>Domestic</td>
<td>5.3222</td>
<td>4.1568</td>
</tr>
<tr>
<td>District</td>
<td>3.2404</td>
<td>1.0318</td>
</tr>
<tr>
<td>percent_aged_under_18_or_over_64</td>
<td>0.2510</td>
<td>0.4597</td>
</tr>
<tr>
<td>percent_aged_16_unemployed</td>
<td>0.2506</td>
<td>1.3331</td>
</tr>
<tr>
<td>percent_of_housing_crowded</td>
<td>0.2016</td>
<td>0.5086</td>
</tr>
<tr>
<td>percent_households_below_poverty</td>
<td>0.2015</td>
<td>1.5177</td>
</tr>
<tr>
<td>percent_aged_25_without_high_sch</td>
<td>0.1479</td>
<td>1.2063</td>
</tr>
<tr>
<td>hardship_index</td>
<td>0.1165</td>
<td>0.8835</td>
</tr>
<tr>
<td>per_capita_income</td>
<td>0.04498</td>
<td>0.9856</td>
</tr>
</tbody>
</table>

Fit Statistics

<table>
<thead>
<tr>
<th>Number of Trees</th>
<th>Training Average Square Error</th>
<th>Validation Average Square Error</th>
<th>Misclassification Rate</th>
<th>Training Log Loss</th>
<th>Validation Log Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1879</td>
<td>0.189</td>
<td>0.2069</td>
<td>0.299</td>
<td>0.506</td>
</tr>
<tr>
<td>2</td>
<td>0.1709</td>
<td>0.174</td>
<td>0.2069</td>
<td>0.299</td>
<td>0.520</td>
</tr>
<tr>
<td>3</td>
<td>0.1572</td>
<td>0.161</td>
<td>0.2069</td>
<td>0.299</td>
<td>0.529</td>
</tr>
<tr>
<td>4</td>
<td>0.1455</td>
<td>0.151</td>
<td>0.2065</td>
<td>0.298</td>
<td>0.492</td>
</tr>
<tr>
<td>5</td>
<td>0.1361</td>
<td>0.143</td>
<td>0.1449</td>
<td>0.156</td>
<td>0.444</td>
</tr>
<tr>
<td>6</td>
<td>0.1279</td>
<td>0.136</td>
<td>0.1264</td>
<td>0.152</td>
<td>0.485</td>
</tr>
<tr>
<td>7</td>
<td>0.1199</td>
<td>0.132</td>
<td>0.1264</td>
<td>0.152</td>
<td>0.485</td>
</tr>
</tbody>
</table>
Perform Logistic Regression with Variable Selection

The `LOGSELECT` procedure fits binary and binomial response models.

```sas
proc logselect data=&dset. noclprint;
  class &target. &nom_input.;
  model &target.(event='1') = &nom_input. &int_input.;
  selection method=stepwise (choose=validate) ;
  partition rolevar=_partind_(train='0' validate='1');
  code file="&outdir./logselect1.sas";
  title "Logistic Regression";
run;

/* The SAS log can include notes for operations on missing values. */
data mycas.ap_scored_logistic;
  set &dset.;
  %include "&outdir./logselect1.sas";
  p_&target.1=p_&target.;
  p_&target.0=1-p_&target.;
run;
```

- The `CLASS` statement names the variables that are used as explanatory variables in the analysis.
- The `MODEL` statement defines the statistical model. The `MODEL` statement and the `CLASS` statement together specify that observations with the formatted value of 1 represent events in the data. The probability that the `LOGSELECT` procedure models is the probability that the variable `arrest_code` takes on the (formatted) value of 1.
- The `SELECTION` statement specifies that the model selection is stepwise regression.
- The `CODE` statement writes SAS DATA step code and saves it in the file `logselect1.sas`. An in-memory DATA step is used with the scoring code.
The LOGSELECT Procedure

Model Information

<table>
<thead>
<tr>
<th>Data Source</th>
<th>CVPART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Variable</td>
<td>arrest_code</td>
</tr>
<tr>
<td>Distribution</td>
<td>Binary</td>
</tr>
<tr>
<td>Link Function</td>
<td>Logit</td>
</tr>
<tr>
<td>Optimization Technique</td>
<td>Newton-Raphson with Ridging</td>
</tr>
</tbody>
</table>

Number of Observations

<table>
<thead>
<tr>
<th>Description</th>
<th>Total</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Observations Read</td>
<td>105951</td>
<td>63971</td>
<td>31786</td>
</tr>
<tr>
<td>Number of Observations Used</td>
<td>85537</td>
<td>57076</td>
<td>28461</td>
</tr>
</tbody>
</table>

Response Profile

<table>
<thead>
<tr>
<th>Ordered Value</th>
<th>arrest_code</th>
<th>Total Frequency</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>60101</td>
<td>90120</td>
<td>19981</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>25436</td>
<td>10955</td>
<td>8480</td>
</tr>
</tbody>
</table>

Probability modeled is arrest_code = 1.

Selection Information

<table>
<thead>
<tr>
<th>Selection Method</th>
<th>Stepwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select Criterion</td>
<td>SBC</td>
</tr>
<tr>
<td>Choose Criterion</td>
<td>Validation ASE</td>
</tr>
<tr>
<td>Stop Criterion</td>
<td>SBC</td>
</tr>
<tr>
<td>Effect Hierarchy Enforced</td>
<td>None</td>
</tr>
<tr>
<td>Stop Horizon</td>
<td>3</td>
</tr>
</tbody>
</table>

Selection Details

Convergence criterion (ABSCCONV=1E-7) satisfied.

Selection Summary

<table>
<thead>
<tr>
<th>Step</th>
<th>Effect Entered</th>
<th>Number Effects In</th>
<th>SBC</th>
<th>Validation ASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Intercept</td>
<td>1</td>
<td>69457.3498</td>
<td>0.2082</td>
</tr>
<tr>
<td>1</td>
<td>111 code</td>
<td>2</td>
<td>45735.2014</td>
<td>0.1251</td>
</tr>
<tr>
<td>2</td>
<td>Location_Description</td>
<td>3</td>
<td>41025.6538</td>
<td>0.1138</td>
</tr>
</tbody>
</table>

Optimal Value Of Criterion

Stepwise selection stopped because adding or removing an effect does not improve the SBC criterion.

The model at step 2 is selected where Validation ASE is 0.113634.

Selected Effects: Intercept 111 code Location_Description

Selected Model

Dimensions

<table>
<thead>
<tr>
<th>Columns in Design</th>
<th>140</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Effects</td>
<td>3</td>
</tr>
<tr>
<td>Max Effect Columns</td>
<td>114</td>
</tr>
<tr>
<td>Rank of Design</td>
<td>133</td>
</tr>
<tr>
<td>Parameters in Optimization</td>
<td>133</td>
</tr>
</tbody>
</table>

Testing Global Null Hypothesis: BETA=0

<table>
<thead>
<tr>
<th>Test</th>
<th>DF</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>132</td>
<td>20662.5642</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Fit Statistics

<table>
<thead>
<tr>
<th>Description</th>
<th>Training</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2 Log Likelihood</td>
<td>40556</td>
<td>20983</td>
</tr>
<tr>
<td>AIC (smaller is better)</td>
<td>40830</td>
<td>21249</td>
</tr>
<tr>
<td>AICC (smaller is better)</td>
<td>40830</td>
<td>21250</td>
</tr>
<tr>
<td>SBC (smaller is better)</td>
<td>42020</td>
<td>22311</td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>1.01084</td>
<td>1.206810</td>
<td>0.7115</td>
<td>0.3989</td>
</tr>
<tr>
<td>111 code Aggravated Assault</td>
<td>1</td>
<td>-2.191491</td>
<td>0.134622</td>
<td>266.0022</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>111 code Aggravated Battery</td>
<td>1</td>
<td>-3.029876</td>
<td>0.131940</td>
<td>527.3459</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>111 code Arson</td>
<td>1</td>
<td>-3.540481</td>
<td>0.344267</td>
<td>151.3209</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
Build a Decision Tree

The **TREESPLIT** procedure is used to build a classification tree.

```plaintext
proc treesplit data=&dset. minleafsize=5 outmodel=mycas.model_treesplit;
    target &target. /level=nominal;
    input &nom_input. /level=nominal;
    input &int_input. /level=interval;
    partition rolevar=_partind_ (train='0' validate='1');
    output out=mycas.ap_scored_treesplit copyvars=(_partind_ &target);
    title "Decision Tree";
run;
```

The TREESPLIT procedure displays a visualization for the first few splits. The visualization is followed by the fit statistics and variable importance.
Assess the Models

Create a Macro for the ASSESS Procedure

- Macros can reduce repetitive code. The macro uses the scored tables that were generated by the analytic procedures (and the in-memory DATA step, in the case of the LOGSELECT procedure).
- The ODS OUTPUT statement is used to create a ROC and lift data set for each model.

```sas
%macro assess_model(prefix=, var_evt=, var_nevt=); PROC ASSESS DATA=MYCAS.AP_SCORED_&prefix.nbins=10; INPUT &var_evt.; TARGET &target. / level=nominal event='1'; FITSTAT PVAR=&var_nevt. / PEVENT='0'; BY _PARTIND_; ODS OUTPUT FITSTAT=WORK.&prefix._FITSTAT ROCINFO=WORK.&prefix._ROCINFO LIFTINFO=WORK.&prefix._LIFTINFO; RUN; %MEND ASSESS_MODEL;
```
title "Assess Gradient Boost";
%assess_model(prefix=gradboost,
    var_evt=p_arrest_code1,
    var_nevt=p_arrest_code0);

title "Assess Logistic Regression";
%assess_model(prefix=logistic,
    var_evt=p_arrest_code1,
    var_nevt=p_arrest_code0);

title "Assess Decision Tree";
%assess_model(prefix=treesplit,
    var_evt=p_arrest_code1,
    var_nevt=p_arrest_code0);

The ASSESS procedure produces result tables for the lift, ROC information, and fit statistics. Instead of showing the tabular results, the data is plotted and shown in the following sections.

Prepare ROC and Lift Data Sets for Plotting

Combine all the ROC information data sets into a single data set. Do the same for the lift information data sets.

data work.all_rocinfo;
  set work.logistic_rocinfo(keep=sensitivity fpr _partind_ in=l)
    work.forest_rocinfo(keep=sensitivity fpr _partind_ in=f)
    work.treesplit_rocinfo(keep=sensitivity fpr _partind_ in=t)
    work.gradboost_rocinfo(keep=sensitivity fpr _partind_ in=g);
  length model $ 16;
  select;
    when (l) model='Logistic';
    when (f) model='Forest';
    when (g) model='GradientBoosting';
    when (t) model='TreeSplit';
  end;
run;

data work.all_liftinfo;
  set work.logistic_liftinfo(keep=depth lift cumlift _partind_ in=l)
    work.forest_liftinfo(keep=depth lift cumlift _partind_ in=f)
    work.treesplit_liftinfo(keep=depth lift cumlift _partind_ in=t)
    work.gradboost_liftinfo(keep=depth lift cumlift _partind_ in=g);
  length model $ 16;
  select;
    when (l) model='Logistic';
    when (f) model='Forest';
    when (g) model='GradientBoosting';
    when (t) model='TreeSplit';
  end;
run;

Plot ROC Curves

ods graphics on;

/* _partind_=2 specifies the test partition */
proc sgplot data=work.all_rocinfo(where=(partind_=2)) aspect=1;
    title "ROC Curves";
    series x=fpr y=sensitivity / group=model;
    lineparm x=0 y=0 slope=1 / transparency=.7;
    yaxis values=(0 to 1 by 0.25) grid offsetmin=.05 offsetmax=.05;
    xaxis values=(0 to 1 by 0.25) grid offsetmin=.05 offsetmax=.05;
run;

- The WHERE clause in the data set option limits the data to the validation data. The training (0) and test (2) partitions are not included in the plot.
- Because the data set includes the ROC information for all the models, the GROUP= option is used to plot a ROC curve for each series independently.
Plot Lift

proc sgplot data=work.all_liftinfo(where=(_partind_=2));
  title "Lift Chart";
  xaxis label="Percentile" grid;
  series x=depth y=lift / group=model markers
      markerattrs=(symbol=circlefilled);
run;

Your lift chart will look similar to the following.

Create Fit Statistics

Similar to the macro that was used for PROC ASSESS, the following code uses a macro to reduce the repetitive code for printing the fit statistics tables.

```
%macro print_fitstats(prefix=);
  proc print data=work.&prefix._fitstat;
  run;
%mend print_fitstats;

title "Forest Fit Statistics";
%print_fitstats(prefix=forest);

title "Gradient Boosting Fit Statistics";
```
```plaintext
%print_fitstats(prefix=gradboost);

title "Logistic Fit Statistics";
%print_fitstats(prefix=logistic);

title "TreeSplit Fit Statistics";
%print_fitstats(prefix=treesplit);

Remember that the partitioning is as follows:

0  training data
1  validation
2  test

### Forest Fit Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>PartInd</em></th>
<th>FormattedPartition</th>
<th>NOBS</th>
<th>ASE</th>
<th>DIV</th>
<th>RASE</th>
<th>MCE</th>
<th>MCLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>63572</td>
<td>0.273192</td>
<td>63572</td>
<td>0.522678</td>
<td>0.164239</td>
<td>0.444120</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>31728</td>
<td>0.275337</td>
<td>31728</td>
<td>0.524726</td>
<td>0.163168</td>
<td>0.447102</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>10651</td>
<td>0.274207</td>
<td>10651</td>
<td>0.523648</td>
<td>0.162614</td>
<td>0.445915</td>
</tr>
</tbody>
</table>

### Gradient Boost Fit Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>PartInd</em></th>
<th>FormattedPartition</th>
<th>NOBS</th>
<th>ASE</th>
<th>DIV</th>
<th>RASE</th>
<th>MCE</th>
<th>MCLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>63572</td>
<td>0.165946</td>
<td>63572</td>
<td>0.407365</td>
<td>0.115356</td>
<td>0.275254</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>31728</td>
<td>0.212294</td>
<td>31728</td>
<td>0.460754</td>
<td>0.138048</td>
<td>0.347726</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>10651</td>
<td>0.213842</td>
<td>10651</td>
<td>0.462430</td>
<td>0.137734</td>
<td>0.349597</td>
</tr>
</tbody>
</table>

### Logistic Fit Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>PartInd</em></th>
<th>FormattedPartition</th>
<th>NOBS</th>
<th>ASE</th>
<th>DIV</th>
<th>RASE</th>
<th>MCE</th>
<th>MCLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>63568</td>
<td>0.225283</td>
<td>63568</td>
<td>0.474646</td>
<td>0.152404</td>
<td>0.363590</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>31724</td>
<td>0.223397</td>
<td>31724</td>
<td>0.472638</td>
<td>0.150643</td>
<td>0.352069</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>10650</td>
<td>0.226001</td>
<td>10650</td>
<td>0.475396</td>
<td>0.150329</td>
<td>0.355661</td>
</tr>
</tbody>
</table>

### TreeSplit Fit Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>PartInd</em></th>
<th>FormattedPartition</th>
<th>NOBS</th>
<th>ASE</th>
<th>DIV</th>
<th>RASE</th>
<th>MCE</th>
<th>MCLL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>63572</td>
<td>0.196722</td>
<td>63572</td>
<td>0.443534</td>
<td>0.131048</td>
<td>0.317611</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>31728</td>
<td>0.219327</td>
<td>31728</td>
<td>0.459663</td>
<td>0.143312</td>
<td>0.532770</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>10651</td>
<td>0.223490</td>
<td>10651</td>
<td>0.472747</td>
<td>0.146841</td>
<td>0.552057</td>
</tr>
</tbody>
</table>
```
Score New Data

About This Section

- This section of the document shows how to get and analyze the latest data from the City of Chicago crime data. As a result, your experience and analysis should be similar, but not identical.

- Much of the code in this section is similar to the code in preceding sections. The differences to notice are as follows:
  - The latest records are downloaded directly from the City of Chicago web service with the JSON engine. A DATA step is used to match the data types from the crimes.sas7bdat file that was used.
  - The FOREST, GRADBOOST, and TREESPLIT procedures were used earlier with an OUTMODEL= option to save their models as in-memory tables. For scoring, the procedures are used again, with an INMODEL= option that reads a model from an in-memory table.
  - The training and validation data used a _PartInd_ column to identify the partition. The latest data does not have this column and it is removed from all the code.

Download and Prepare the Latest Data

- The first two lines of code cause the SAS Workspace Server to download the most recent 1000 records.
- The first DATA step sets the data types and formats to match the existing data.
- The second DATA step joins the most recent records with the census data. The date that is specified in the WHERE clause ensures that the records have not been processed already. This results in far fewer than 1000 records.

```sas
filename chicago url 'https://data.cityofchicago.org/resource/6zsd-86xi.json';
libname chicago sasejson;

data mycas.arrest err;
set chicago.root(rename=(
  date=tmpts arrest=arrest_code domestic=tmpds updated_on=tmpuo
  beat=tbeat community_area=tca district=tdis id=tid
  longitude=tlong latitude=tlat ward=tward
  x_coordinate=tx y_coordinate=ty year=tyear
  ));

if arrest_code eq 0 then arrest = 'false';
else arrest = 'true';
if tmpds eq 0 then domestic = 'false';
else domestic = 'true';

  beat = input(tbeat, best12.);
  community_area = input(tca, best12.);
  district = input(tdis, best12.);
  id = input(tid, best12.);
  longitude = input(tlong, best12.);
  latitude = input(tlat, best12.);
  ward = input(tward, best12.);
  x_coordinate = input(tx, best12.);
  y_coordinate = input(ty, best12.);
```

27
year = input(tyear, best12.);

format fbi_code $fbi. block $35. description $59. location_description $47. 
   primary_type $33. arrest domestic $5. 
   arrest_code 8. date mmddyy10. timestamp updated_on datetime. ;

pos = kindex(tmpts, 'T');
if -1 eq pos then output err;
date = input(substr(tmpts,1,pos-1), yymmdd10.);
time = input(substr(tmpts,pos+1), time.);
timestamp = dhms(date,0,0,time);

pos = kindex(tmpuo, 'T');
if -1 eq pos then output err;
uodate = input(substr(tmpuo,1,pos-1), yymmdd10.);
uotime = input(substr(tmpuo,pos+1), time.);
updated_on = dhms(uodate,0,0,uotime);

drop tmpts tmpuo pos time tmpds uodate uotime tbeat tca tdis tid 
   tlong tlat tward tx ty tyear;
output mycas.arrest;
run;

data mycas.latest_crimes;
   merge 
      mycas.arrest(in=in1 where=(date > '15FEB2016'd))
      mycas.census(in=in2);
   by community_area;
   if in1;
run;

Note: If you receive an error in the SAS log that is related to missing a CA trust list, then locate the trustedcerts.pem file that is part of the SAS installation, submit code like the following, and then rerun:

   options sslcalistloc="/path/to/trustedcerts.pem";

Score the Latest Data

data mycas.latest_logistic;
   set mycas.latest_crimes;
   %include "/&outdir./logselect1.sas";
   p_&target.1=p_&target.;
   p_&target.0=1-p_&target.;
run;

proc treesplit data=mycas.latest_crimes inmodel=mycas.model_treesplit noprint;
   target &target. /level=nominal;
   input &nom_input. /level=nominal;
   input &int_input. /level=interval;
   output out=mycas.latest_treesplit copyvars=(&target);
run;

proc gradboost data=mycas.latest_crimes inmodel=mycas.model_gradboost noprint;
   output out=mycas.latest_gradboost copyvars=(&target);
run;
An in-memory DATA step is used to score the data with the model from PROC LOGSELECT. The other procedures use the INMODEL= option to read the model from an in-memory table.

The result of this code is an additional four in-memory tables. You can use the ASSESS procedure again to see how the models operated on the latest data.

Assess the Models on the Latest Data

%macro assess_latest(prefix=, var_evt=, var_nevt=);
   proc assess data=mycas.latest_&prefix. nbins=10;
      input &var_evt.;
      target &target. / level=nominal event='1';
      fitstat pvar=&var_nevt. / pevent='0';
      ods output fitstat=work.&prefix._lfitstat
         rocinfo=work.&prefix._lrocinfo
         liftinfo=work.&prefix._lliftinfo;
   run;
%mend assess_latest;

title "Assess Decision Tree";
%assess_latest(prefix=treesplit, var_evt=p_arrest_code1, var_nevt=p_arrest_code0);

title "Assess Forest";
%assess_latest(prefix=forest, var_evt=p_arrest_code1, var_nevt=p_arrest_code0);

title "Assess Gradient Boost";
%assess_latest(prefix=gradboost, var_evt=p_arrest_code1, var_nevt=p_arrest_code0);

title "Assess Logistic Regression";
%assess_latest(prefix=logistic, var_evt=p_arrest_code1, var_nevt=p_arrest_code0);

Combine the ROC and lift data sets that were written to the Work libref the SAS client.

data work.latest_rocinfo;
   set work.logistic_lrocinfo(keep=sensitivity fpr in=l)
      work.forest_lrocinfo(keep=sensitivity fpr in=f)
      work.treesplit_lrocinfo(keep=sensitivity fpr in=t)
      work.gradboost_lrocinfo(keep=sensitivity fpr in=g);
   length model $ 16;
   select;
      when (l) model='Logistic';
      when (f) model='Forest';
      when (g) model='GradientBoosting';
      when (t) model='TreeSplit';
   end;
run;

data work.latest_liftinfo;
   set work.logistic_lliftinfo(keep=depth lift cumlift in=l)
      work.forest_lliftinfo(keep=depth lift cumlift in=f)
      work.treesplit_lliftinfo(keep=depth lift cumlift in=t)
      work.gradboost_lliftinfo(keep=depth lift cumlift in=g);
   length model $ 16;
select;
when (l) model='Logistic';
when (f) model='Forest';
when (g) model='GradientBoosting';
when (t) model='TreeSplit';
end;
run;

Plot ROC Curves and Lift Chart

proc sgplot data=work.latest_rocinfo aspect=1;
  title "ROC Curves";
  title2 "Latest Data";
  series x=fpr y=sensitivity / group=model;
  lineparm x=0 y=0 slope=1 / transparency=.7;
  yaxis values=(0 to 1 by 0.25) grid offsetmin=.05 offsetmax=.05;
  xaxis values=(0 to 1 by 0.25) grid offsetmin=.05 offsetmax=.05;
run;

proc sgplot data=work.latest_liftinfo;
  title "Lift Chart";
  title2 "Latest Data";
  xaxis label="Percentile" grid;
  series x=depth y=lift / group=model markers markerattrs=(symbol=circlefilled);
run;
Additional Reading

Here is the recommended reading list for this title:

- SAS Cloud Analytic Services: Accessing and Manipulating Data
- SAS Cloud Analytic Services: Fundamentals
- SAS Cloud Analytic Services: Language Reference