Viya Questions for Model Validation

How to use AIC criterion for variable selection

- AIC is a penalized fit measure used to help reduce overfitting by penalizing a model for the number of parameters.
- Is applied to training data
- Can be used to select variables within a Generalized Linear Model (e.g. Linear or Logistic Regression), in combination with other selection criteria such as validation error.
- AIC = -2LL + 2p, i.e. the negative 2 Log Likelihood + 2 x the number of model parameters
  - Smaller is better
- SBC is a similar penalized fit measure that tends to give you smaller models. SBC = -2LL + plog(n), i.e. the negative 2 Log Likelihood + the number of model parameters x the log of sample size in training data.
  - Smaller is better
- AIC and SBC are relative measures. Results are not comparable between different data sets.

Variable Reduction on SAS Viya

- SAS Viya offers both Supervised & Unsupervised Variable Reduction methods in the Visual and Programmatic interfaces
- Supervised methods find variables that are related to the outcome
- Unsupervised methods finds variables that explain most variance in the input space and reduce correlation between inputs.
- Variable Clustering is also available with VDMML (PROC GVARCLUS), and in SAS 9 (PROC VARCLUS), in which variables that are correlated are placed in the same cluster. Best representative variables can be selected from each cluster.

Variable Reduction and Variable Clustering on SAS Viya

- When PROC VARREDUCE (Variable Selection/Reduction)performs **unsupervised** variable reduction, it analyzes variance and reduces dimensionality by forward selection of the variables that contribute the most to the overall data variance.
- When PROC VARREDUCE performs supervised variable selection, it analyzes the variance and reduces dimensionality by forward selection of the variables that contribute the most to explaining the overall variance of the *response variables* (targets). Supervised selection also optionally includes Decision Tree based (DT, RF, GB) selection. **All methods can be combined!**
- Variable clustering (PROC GVARCLUS) divides numeric variables into disjoint or hierarchical clusters. Variables in different clusters are conditionally independent given their own clusters. The clustering process starts with one variable per cluster, and the number of variables per cluster increases as clusters are merged. The variable that contributes the most to the variation in that cluster is chosen as the representative variable.

SAS 9 on Viya

Viya has two compute engines:

- SAS Programming Runtime Environment (SPRE)
  - Allows you to run SAS 9 code and STAT procedures such as LOGISTIC, GENMOD, ARIMA etc.
  - License of VDMML gives you access to BASE SAS and SAS/STAT
  - License of Econometrics or Visual Forecast gives you access to SAS/ETS

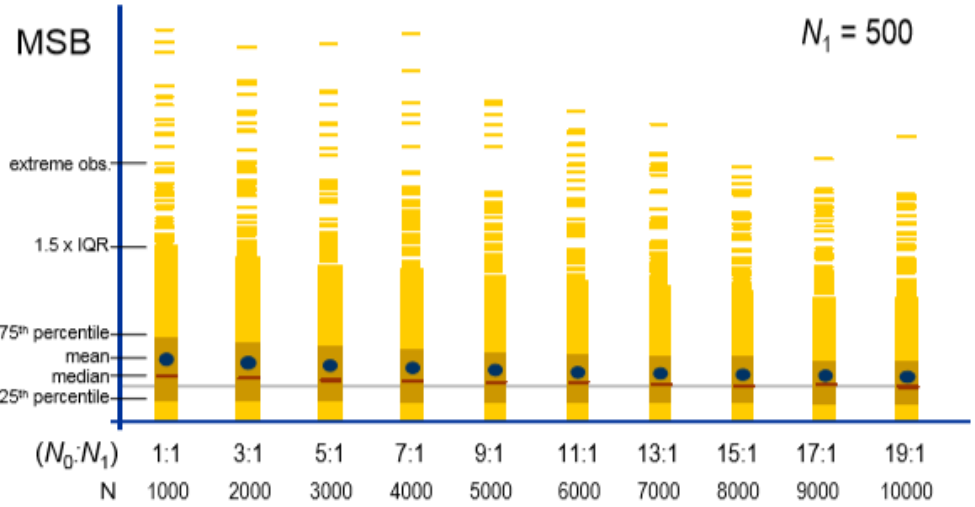
- Cloud Analytic Service (CAS), the in-memory engine
  - Data is distributed CAS worker nodes and threads.
  - Processing is done in parallel

Oversampling

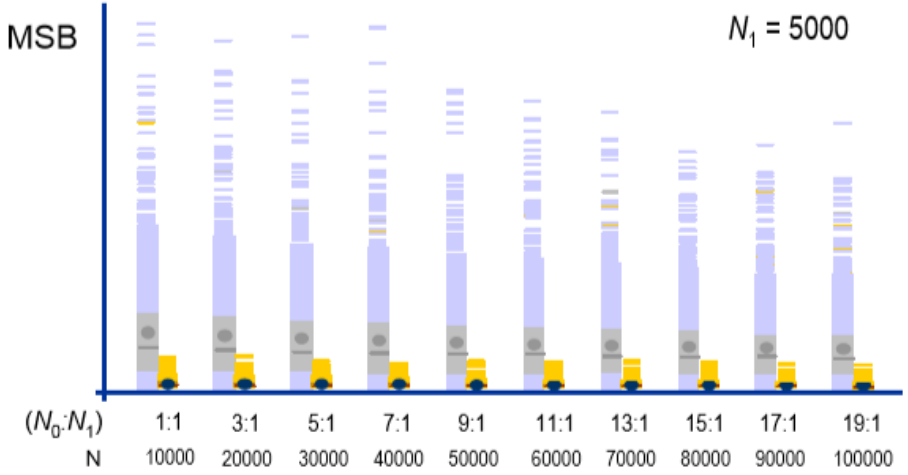
- Oversampling increases the efficiency of modeling by reducing the overall size of the modeling data set and provides similar model performance. It results in similar predictive performance compared to non-oversampled data.
- Oversampling and posterior probability adjustments are available in Model Studio
- There is no package to optimize oversampling rates. You can run your own simulation or use trial and error.
- Typically, I use oversampling for percentages under 10 or 5%, and oversample to about 30 to 40%. For a small number of target events, 50% may discard too much data.

$$p_1 = \frac{\tilde{p}_1(\pi_1/\rho_1)}{\tilde{p}_0(\pi_0/\rho_0) + \tilde{p}_1(\pi_1/\rho_1)}$$

Separate Sampling Simulation Results



Simulation Results for Larger Sample



SAS Package to convert Categorical to numeric variables?

SAS models automatically accommodate categorical variables.

- In parametric models such as regression and neural networks they are automatically recoded to numeric flags.
- For Tree based methods (decision trees, GB, RF) categorical variables are accommodated directly through the algorithm.
  - E.g. ”For a categorical variable split search, how many ways can I break a k-level unordered variable into n groups?”
- No pre-processing or package is required. This is a great advantage of SAS.

Conversion of Categorical Variables to WOE or IV can be performed using Data Step programming.

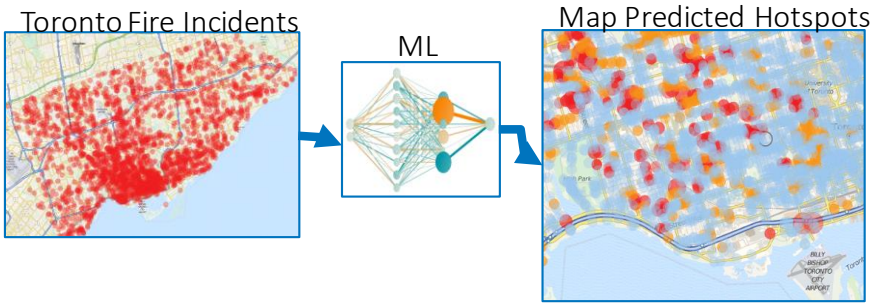
- Capability will be available in the Credit Risk Solution.

SAS Studio on Viya

sas.com

How do you explain the importance and impact of drivers?

Model Interpretation



- Profile Model Prediction Segments
- GeoMap
- Surrogate Model
- Variable Importance using Decision Tree based models (DT, GB, RF)
- Partial Dependence
- Individual Conditional Expectation
- Local Interpretable Model-agnostic Explanations & Shapley

How do you explain the importance and impact of drivers?

Explain Model Action Set

Explain Model Actions Action Set: Syntax

CASL | Lua | Python | R

Provides actions for explaining already trained models.

Syntax ▾ Details ▾ Examples ▾

Table of Actions

Action Name	Description
linearExplainer	Uses linear models to explain already trained models. Supports global linear surrogates as well as the local methods: LIME and KERNEL SHAP.
partialDependence	Computes the partial dependence of an already trained model.

Shapely example:

https://documentation.sas.com/?docsetId=casactml&docsetVersion=8.4&docsetTarget=casactml_explainmodel_examples06.htm&locale=en

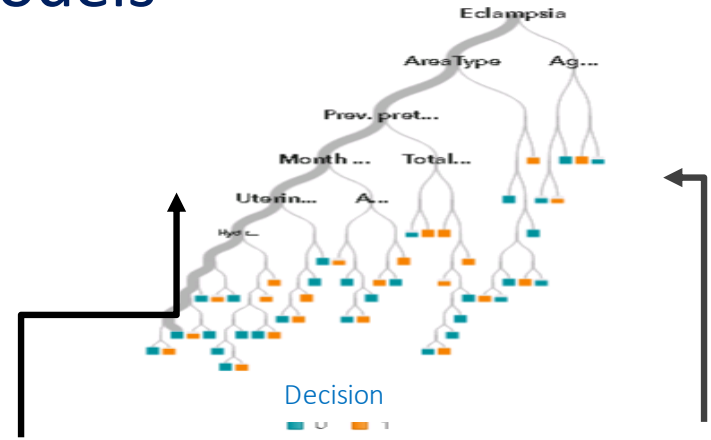
Surrogate Models

Approximate & Explain:

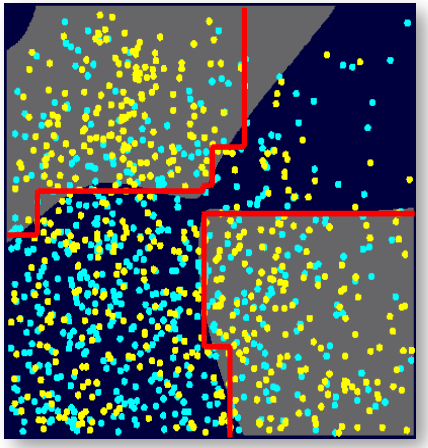
- Supervised ML predictions or decisions
- Unsupervised classifications

… using a simple ‘Surrogate’

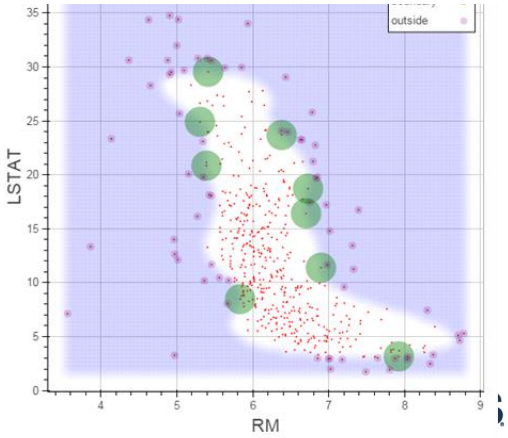
Surrogate Decision Tree Model



Supervised Neural Network

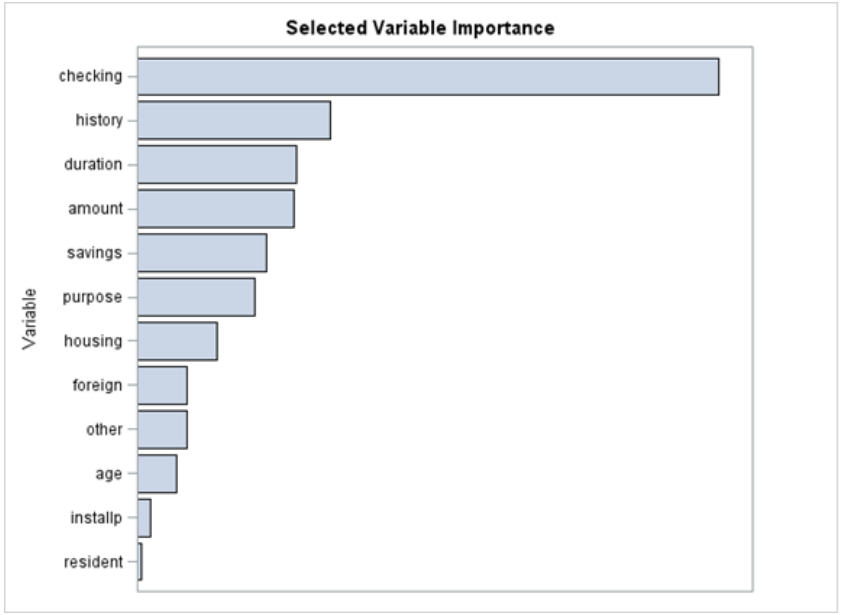


Unsupervised Anomaly Detector: SVDD

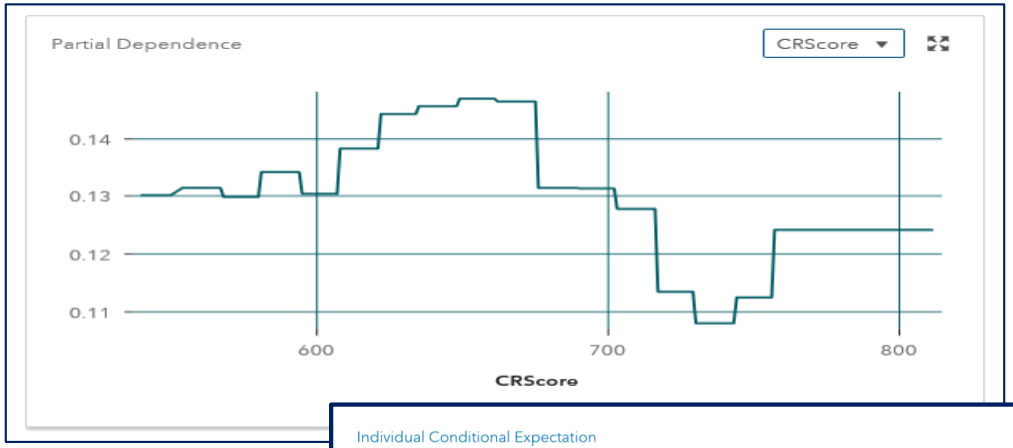


Variable Importance

- A relative measure of an input's contribution to impurity reduction or correct prediction
- Does not provide information on direction of relationships between inputs and target
- A feature of Decision Tree, Random Forest, Gradient Boosting algorithms
- Tree-methods can be fit to predictions from other algorithms (e.g. neural networks, SVMs) to provide a 'surrogate' variable importance



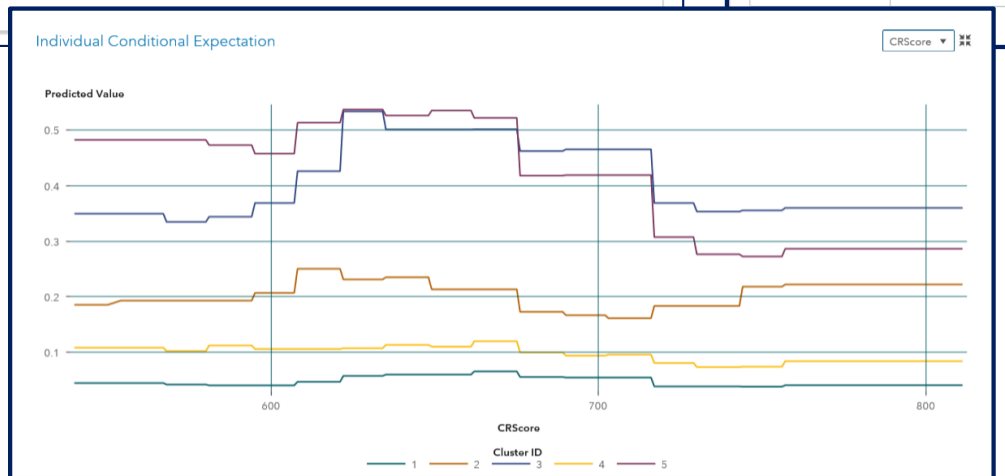
Model Interpretability: PD, ICE, LIME



Local Model

Cluster Centroid: 1

Effect	Parameter	Degrees o...	Estimate	Standard...
Intercept	Intercept	1	0.2294	0.0036
SavBal	SavBal	1	0.0000	0.0000
Sav	Sav 0	1	-0.0614	0.0025
Sav	Sav 1	0	0	.
CD	CD 0	1	-0.0859	0.0035
CD	CD 1	0	0	.
DDABal	DDABal	1	0.0000	0.0000
ATMAmt	ATMAmt	1	0.0000	0.0000



Additional Links

Variable Selection (visual)
https://go.documentation.sas.com/?cdcId=capcdc&cdcVersion=8.4&docsetId=vdmmlref&docsetTarget=n195h0sj6149mfn1y1rain1v19fz.htm&locale=en&activeCdc=vdmmlcdc
Variable Clustering (visual)
https://go.documentation.sas.com/?cdcId=capcdc&cdcVersion=8.4&docsetId=vdmmlref&docsetTarget=n03yg60d2z6gobn1unqbo4w4q7h2.htm&locale=en&activeCdc=vdmmlcdc&docsetVersion=8.4

PROC VARREDUCE
http://documentation.sas.com/?docsetId=casstat&docsetVersion=8.4&docsetTarget=casstat_varreduce_toc.htm&locale=en
PROC GVARCLUS
https://documentation.sas.com/?docsetId=casml&docsetTarget=casml_gvarclus_toc.htm&docsetVersion=8.4&locale=en

Machine Learning Interpretability including LIME, ICE, SHAP
https://go.documentation.sas.com/?cdcId=pgmcdc&cdcVersion=8.11&docsetId=casactml&docsetTarget=casactml_explainmodel_toc.htm&locale=en
https://go.documentation.sas.com/?cdcId=capcdc&cdcVersion=8.4&docsetId=vdmmladvug&docsetTarget=n0dqb7uyhlqogvn16xnengpro0fc.htm&locale=en&activeCdc=vdmmlcdc

Distribution Fitting CDF (Returns a value from a cumulative probability distribution)
https://go.documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=lefunctionsref&docsetTarget=n0n7cce4a3gfqkn1vr0p1x0of99s.htm&locale=en
PROC UNIVARIATE distribution fitting
http://support.sas.com/documentation/cdl/en/procstat/67528/HTML/default/viewer.htm#procstat_univariate_examples22.htm
Discrete Distributions (Poisson)
https://blogs.sas.com/content/iml/2012/04/04/fitting-a-poisson-distribution-to-data-in-sas.html

Model Studio on Viya

sas.com

Q&A

sas.com