Evaluating Predictive Models

Comparing Models from SSAS, R, Python, and Azure ML

Dejan Sarka

Our Partners

Quest

DBPLUS better performance

IDERA

SolidQ Think Big. Move Fast.

trivadis makes IT easier.

SOLISYON

PASS

Microsoft

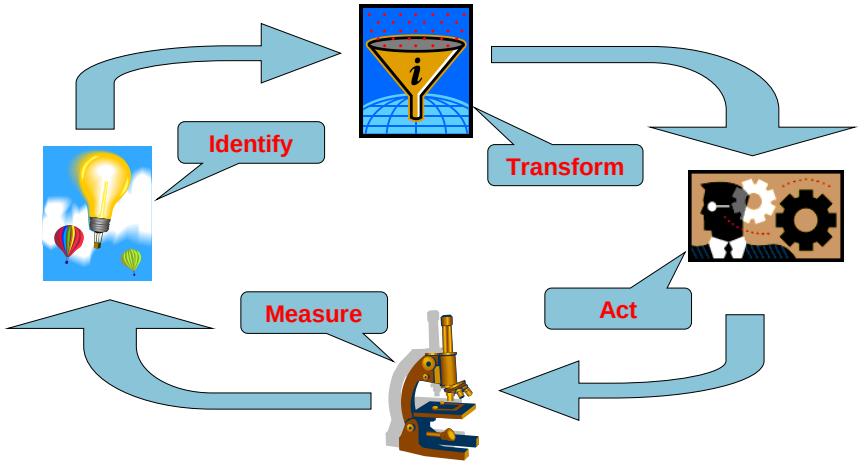
Introduction

- Dejan Sarka (dsarka@solidq.com, dsarka@siol.net, @DejanSarka)
  - 30 years of experience
  - SQL Server MVP, MCT,…
  - 16 books
  - 15+ courses
  - Focus:
    - Data modeling
    - Data mining
    - Data quality

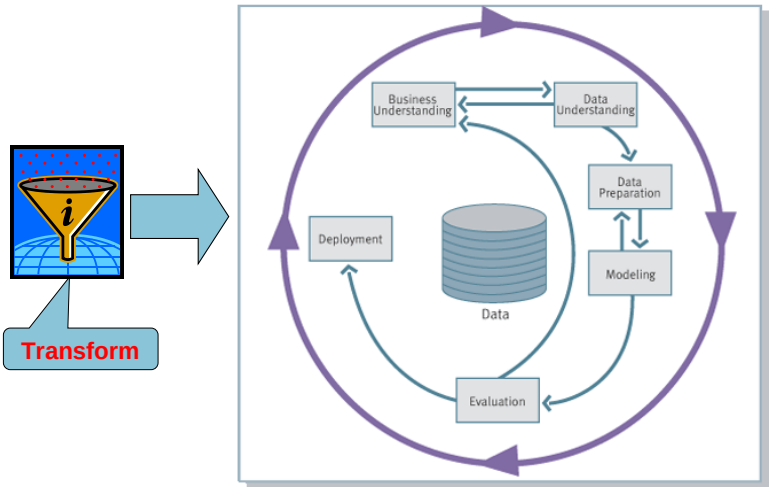
Agenda

- Advanced analytics virtual cycle
- Training and test sets
- Evaluating predictive models
- Classification (confusion) matrix and derivatives
- Evaluating in SSAS, R, Python, and Azure ML models
- Bringing it all together

Advanced Analysis Virtuous Cycle



The CRISP Model



CRISP = Cross Industry Standard Process for Data Mining
(http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining)

Training and Test Sets

- For predictive models, you need to split the data into training and test sets in order to evaluate the models
  - A *training* set is required to build the model (70% of the data)
  - A *test* set is used for predictions (30% of the data)
  - When you know the value of the predicted variable, you can measure the quality of the predictions
- As with every sampling, it is important to *randomly* select the data for each set

Testing the Sampling

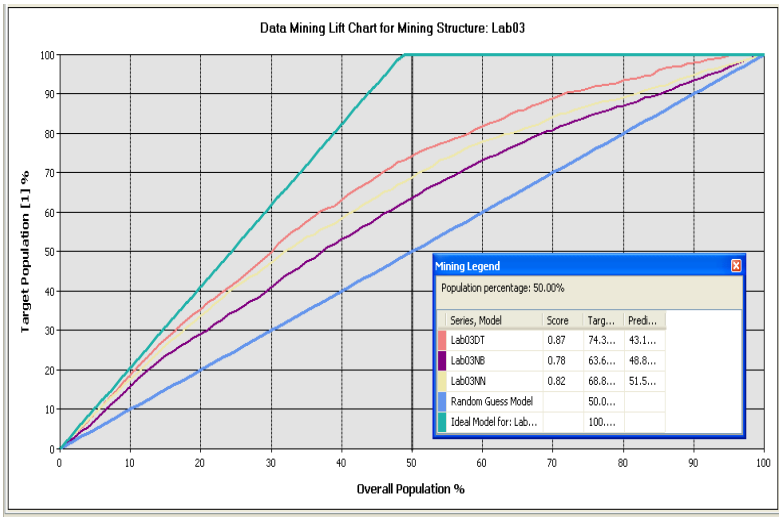
- Test for randomness by creating the two samples and adding a variable that identifies whether a set is a training set or test set (or that identifies the sample)
- Check the null hypothesis that the new variable is not related to other variables
  - Use one-way analysis of variance (ANOVA) and F-tests for continuous variables
  - Use chi-squared test for discrete variables
  - Can also use Decision Trees for discrete variables

Evaluating Predictive Models

- Lift chart
- Cross validation
- Classification (confusion) matrix and derivatives

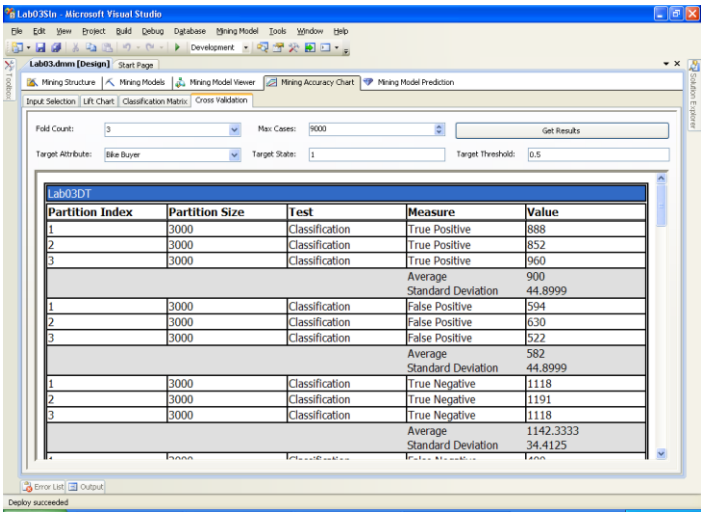
Lift Chart

- No target value: overall performance
- Target value: a percentage of the target audience against a specified percentage of the complete audience



Cross Validation

- Cross validation show robustness of models
  - Splits training set in folds
  - Use one fold for testing, others for training
  - You can see how models perform over different subsets of data



Classification (Confusion) Matrix

- Columns represent items that have been predicted
- Rows represent the actual state of the attribute

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

Derivatives from Confusion Matrix

- Sensitivity, recall, hit rate, or true positive rate (TPR)

$$\frac{TP}{TP + FN}$$

- Specificity or true negative rate (TNR)

$$\frac{TN}{TN + FP}$$

- Precision or positive predictive value (PPV)

$$\frac{TP}{TP + FP}$$

Derivatives from Confusion Matrix

- Negative predictive value (NPV)

$$\frac{TN}{TN + FN}$$

- Miss rate or false negative rate (FNR)

$$\frac{FN}{FN + TP}$$

- Fall-out or false positive rate (FPR)

$$\frac{FP}{FP + TN}$$

Derivatives from Confusion Matrix

- False discovery rate (FDR)

$$\frac{FP}{FP+TP}$$

- False omission rate (FOR)

$$\frac{FN}{FN+TN}$$

- Accuracy (ACC)

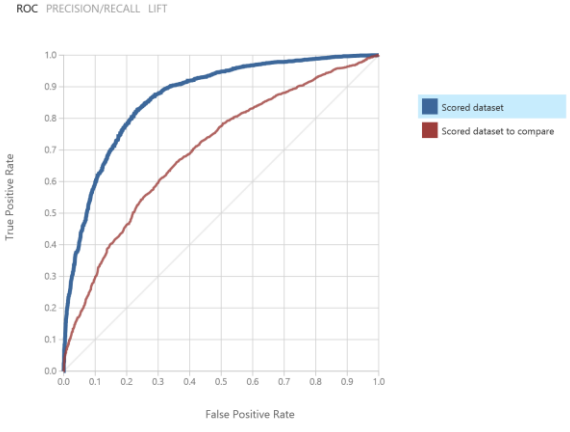
$$\frac{TP+TN}{TP+TN+FP+FN}$$

Derivatives from Confusion Matrix

- Harmonic mean of precision and sensitivity (F1 score)

$$\frac{2*TP}{2*TP+FP+FN}$$

- Receiver operating characteristic (ROC) curve
  - Plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings



Evaluating in SSAS, R, and Azure ML

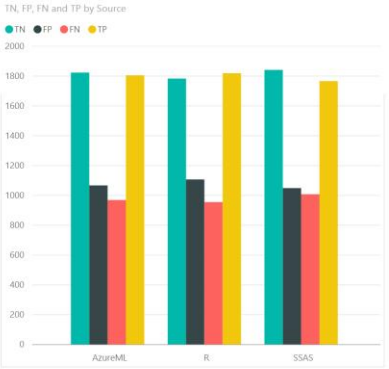
- SSAS
  - Lift chart
  - Cross-validation
  - Classification matrix (basic)
- R, Python
  - Whatever you code☺
- Azure ML
  - ROC
  - Classification matrix (partly)

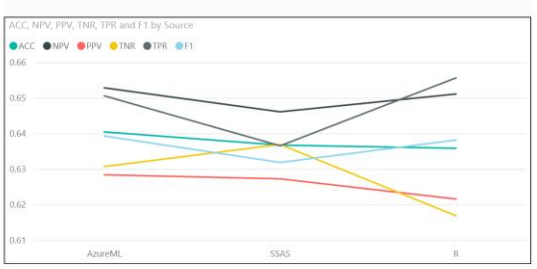
Bringing It All Together

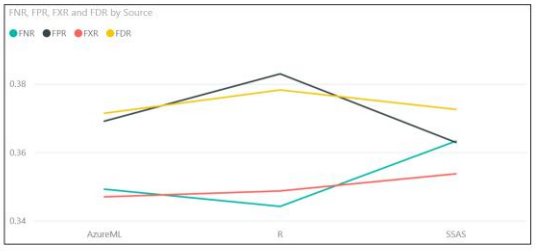
- Use DMX (data mining extensions) query to browse SSAS mining models and perform predictions
- Use T-SQL to execute R and Python code and return tabular format
  - sys. sp_execute_external_script (SQL Server 2016 and 2017)
  - PREDICT() function (2017 only, limited usability)
- Use Azure ML Web services to get predictions from Azure ML models in your application or in Excel
- Use SSIS to union all predictions and create a report

Bringing It All Together

Source	TN	FP	FN	TP
AzureML	1823	1067	969	1805
R	1783	1107	955	1819
SSAS	1841	1049	1008	1766







Q & A

- Thank you!