Автоматизация исследований и формирование аналитических проектов из ad-hoc исследований

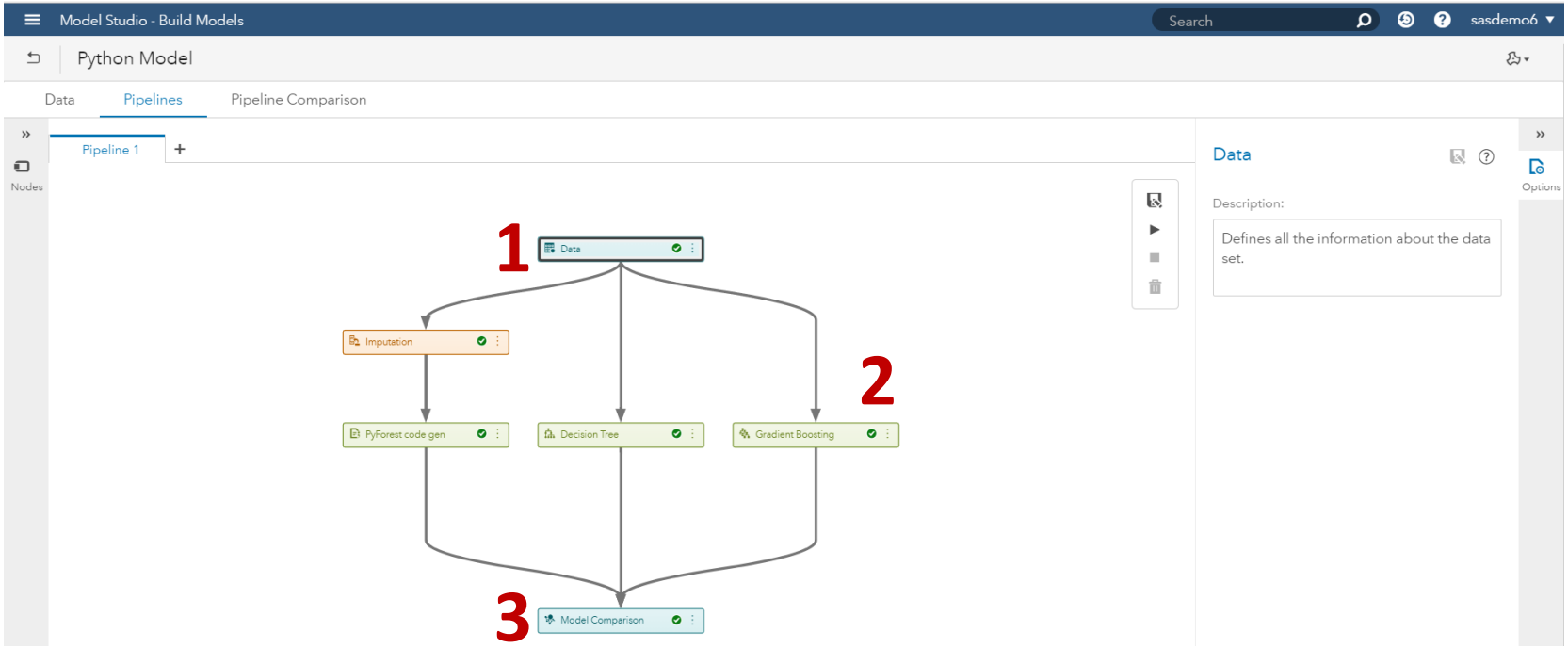
Дмитрий Онищенко

Старший консультант, Дирекция аналитических и индустриальных решений, SAS Россия / СНГ

Copyright © SAS Institute Inc. All rights reserved.

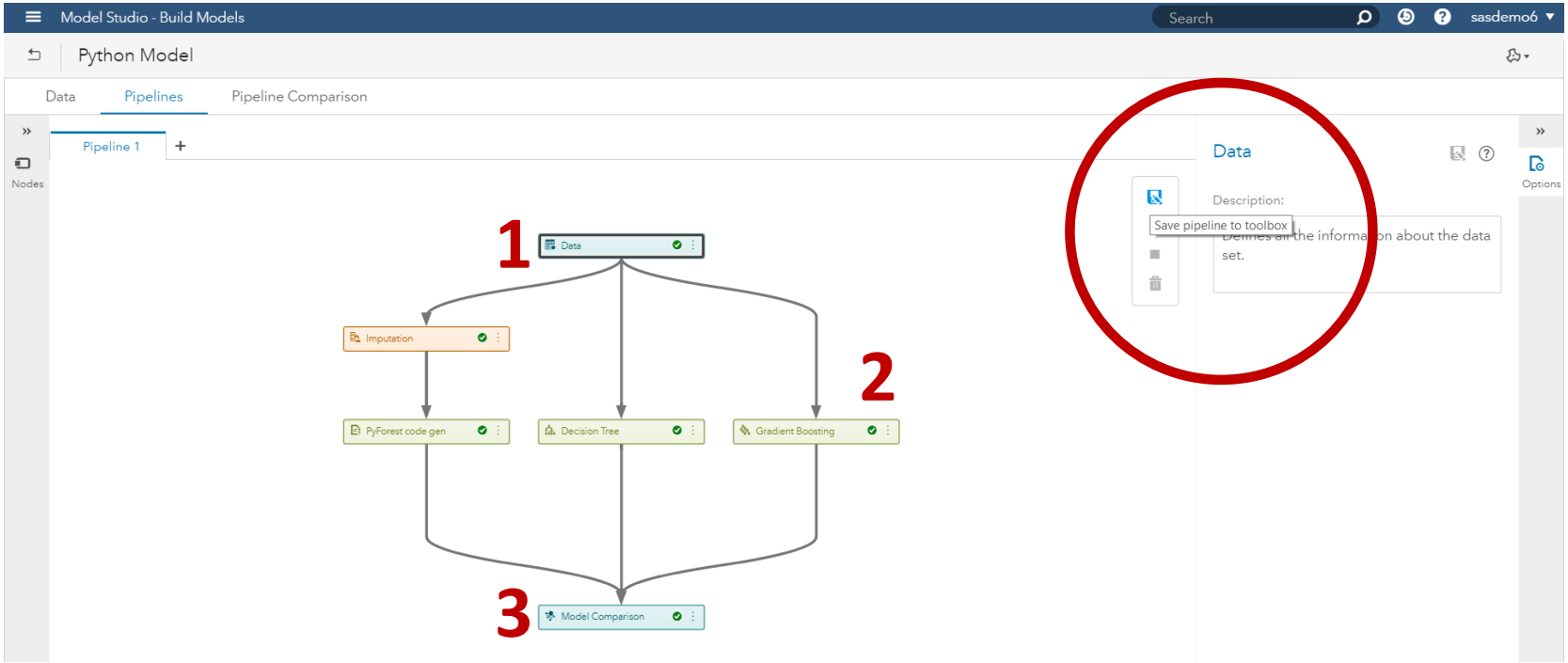
THE POWER TO KNOW.

SAS Model Studio и шаблоны моделей



Copyright © SAS Institute Inc. All rights reserved.

SAS Model Studio и шаблоны моделей



Copyright © SAS Institute Inc. All rights reserved.

Автоматически строим и обновляем модели

Model Studio - Build Models

Search

sasdemo6

Python Model

Data | Pipelines | Pipeline Comparison

Filter

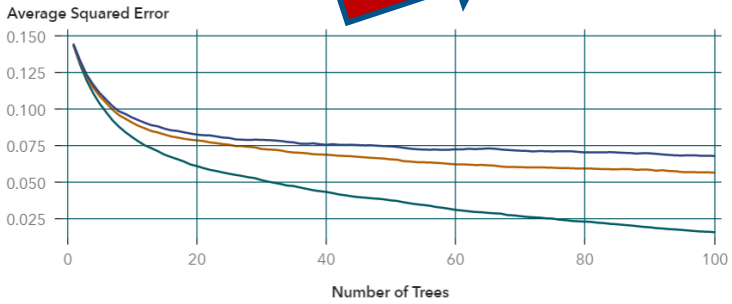
Data: Validate

Compare

Champion	Name	Algorithm Name	KS (Youden)	Pipeline Name
	Gradient Boosting	Gradient Boosting	0.807	Pipeline 1

Error Plot

Average Squared Error



Variable Importance

Variable Name	Train Importance	Importance Standa...	Relative Importance
DEBTINC	16.5764	25.8308	1
DELINQ	5.3890	2.9675	0.3251
NINQ	5.1637	1.0954	0.3115
YOJ	4.7390	1.3971	0.2859
CLNO	4.6769	0.9758	0.2821
JOB	4.3015	1.0827	0.2595
CLAGE	3.9411	1.7921	0.2378
VALUE	3.6378	2.0029	0.2195

Copyright © SAS Institute Inc. All rights reserved.

Демонстрация

Copyright © SAS Institute Inc. All rights reserved.

Промежуточный итог: главные шаги

Ad-hoc модели

Pipeline (конвейер машинного обучения)
Развитие + кастомизация

Сохранение шаблонов конвейеров
(best practices templates)

Сохранение моделей:

- в CAS
- в Model Manager

Copyright © SAS Institute Inc. All rights reserved.

Претендентов много – чемпион один

Model Studio - Build Models

Search

sasdemo6

SAS Demo Day Complete

Data | Pipelines | Pipeline Comparison

Filter

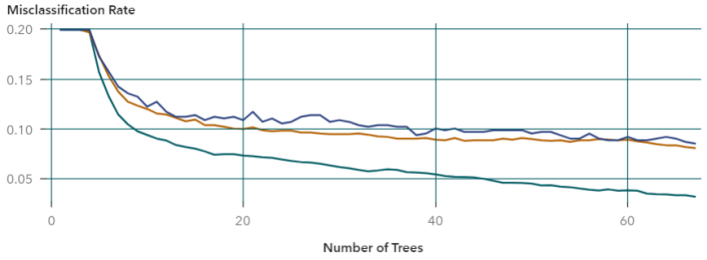
Data: Validate

Compare

Champion	Name	Algorithm Name	Pipeline Na...	KS (Youden)
☑	Gradient Boosting	Gradient Boosting	Pipeline 1	0.799
☐	Gradient Boosting	Gradient Boosting	Pipeline 2	0.799
☐	Gradient Boosting	Gradient Boosting	Pipeline 3	0.798
☐	Forest	Forest	Pipeline 4	0.726

Error Plot

Misclassification Rate



Variable Importance

Variable Name	Train Importance	Importance Standard ...	Relative Importance
DEBTINC	26.1116	34.3086	1
DELINQ	7.5901	3.2059	0.2907
NINQ	6.9218	1.3102	0.2651
JOB	5.8431	1.3393	0.2238
YOJ	5.6211	1.1884	0.2153
CLNO	5.1238	1.2805	0.1962
CLAGE	4.9419	1.8588	0.1893
MORTDUE	4.4407	1.5535	0.1701

Copyright © SAS Institute Inc. All rights reserved.

А судьи кто?
Диагностика и сравнение

Model Studio - Build Models

Search

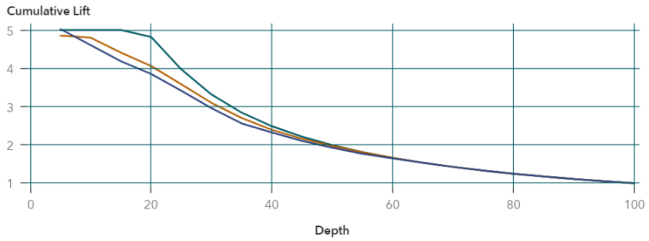
sasdemo6

Python Model

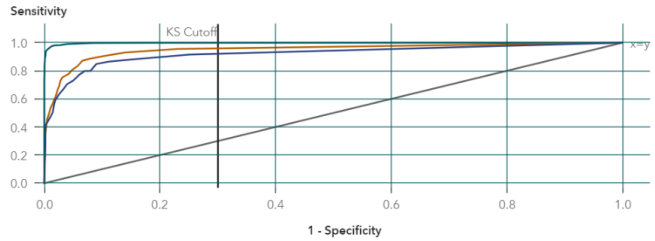
1

Data | Pipelines | Pipeline Comparison

Lift Reports — Cumulative Lift



ROC Reports — ROC



Fit Statistics

2

Data Role	Partition Indicator	Formatted Partiti...	Sum of Frequen...	Average Square...
TEST	2	2	596	0.0682
TRAIN	1	1	3,576	0.0160
VALIDATE	0	0	1,788	0.0567

Train Code

3

```
*-----------------------------------------------------------*;
* Macro Variables for input, output data and files;
  %let dm_datalib =;
  %let dm_lib     = WORK;
  %let dm_folder  = %sysfunc(pathname(work));
*-----------------------------------------------------------*;
*-----------------------------------------------------------*;
  * Training for gradboost;
*-----------------------------------------------------------*;
proc gradboost data=&dm_datalib..DM_F2PC86S6CO1X7AA23WST49EG(&dm_data_caslib)
     earlystop(tolerance=0 stagnation=5)
     numBin=20
     maxdepth=6
```

Copyright © SAS Institute Inc. All rights reserved.

Предикторы на вход модели по-прежнему можно отобрать вручную

Model Studio - Build Models

Search

sasdemo6

Python Model

Data | Pipelines | Pipeline Comparison

Score Inputs

Uppercased Name	Name	Role	Variable Level	Type
CLAGE	CLAGE	INPUT	INTERVAL	N
CLNO	CLNO	INPUT	INTERVAL	N
DEBTINC	DEBTINC	INPUT	INTERVAL	N
DELINQ	DELINQ	INPUT	NOMINAL	N
DEROG	DEROG	INPUT	NOMINAL	N
JOB	JOB	INPUT	NOMINAL	C
LOAN	LOAN	INPUT	INTERVAL	N
MORTDUE	MORTDUE	INPUT	INTERVAL	N

Score Outputs

Name	Uppercased Name	Role	Type	Variable Type
EM_CLASSIFICATION	EM_CLASSIFICATION	CLASSIFICATION	C	
EM_EVENTPROBABILITY	EM_EVENTPROBABILITY	PREDICT	N	
EM_PROBABILITY	EM_PROBABILITY	PREDICT	N	
I_BAD	I_BAD	CLASSIFICATION	C	
P_BAD0	P_BAD0	PREDICT	N	
P_BAD1	P_BAD1	PREDICT	N	
WARN	_WARN_	ASSESS	C	

EP Score Code

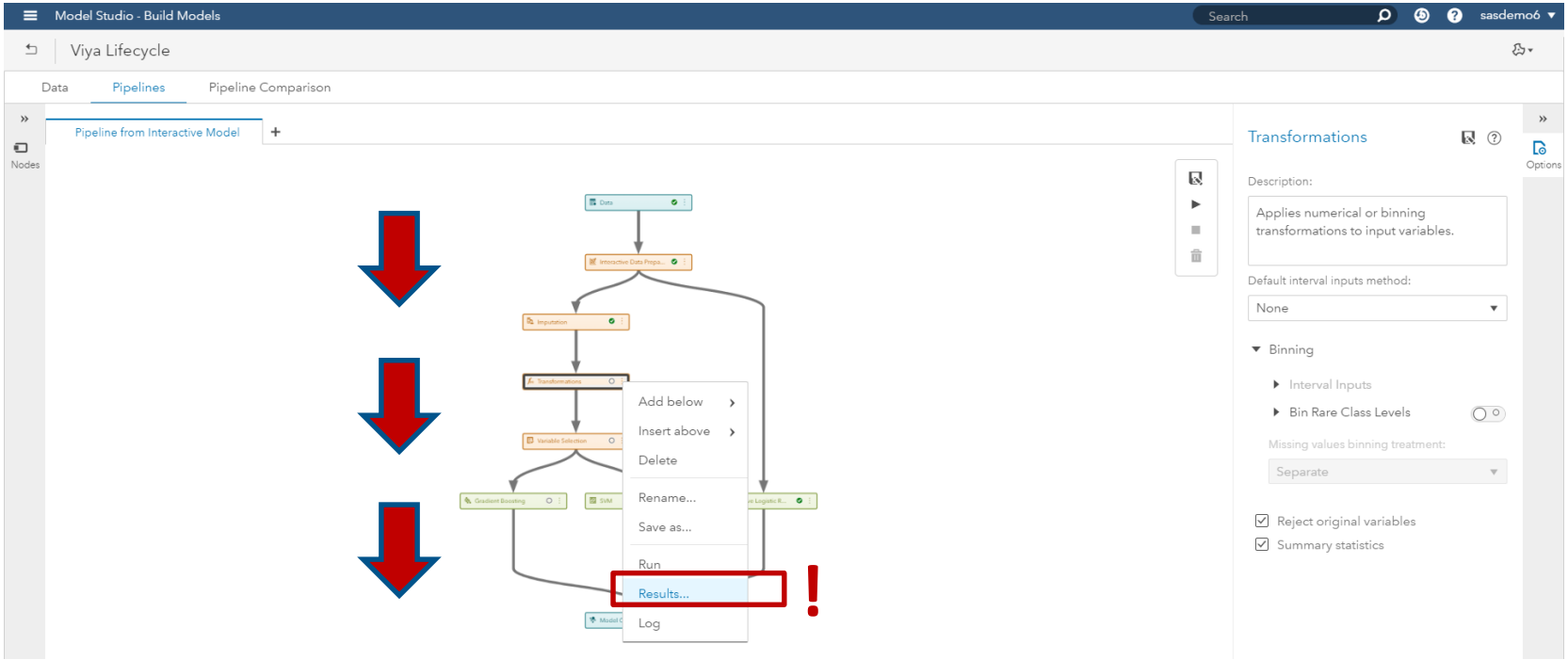
```
data sasep.out;
    dcl package score _9ZE5XBMAAGMFEX3XYN3GMGBQS();
    dcl double "CLAGE";
    dcl double "CLNO";
    dcl double "DEBTINC";
    dcl double "LOAN";
    dcl double "MORTDUE";
    dcl double "VALUE";
    dcl double "YOJ";
    dcl double "DELINQ";
    dcl double "DEROG";
    dcl double "NINQ";
    dcl nchar(14) "REASON";
```

Input Relative Importance

Name	Relative Importance	Variable Level	Variable Label
CLAGE	0.0942	INTERVAL	
CLNO	0.0913	INTERVAL	
DEBTINC	1	INTERVAL	
DELINQ	0.3699	NOMINAL	
DEROG	0.2332	NOMINAL	
JOB	0.0449	NOMINAL	
LOAN	0.1279	INTERVAL	
MORTDUE	0.0502	INTERVAL	

Copyright © SAS Institute Inc. All rights reserved.

Промышленные модели: построение и результаты в реальном времени



Copyright © SAS Institute Inc. All rights reserved.

Промышленные модели: построение и результаты в реальном времени

Model Studio - Build Models

Search

sasdemo6

Viya Lifecycle › Imputation Results

Close

Input Variable Statistics

Input Variable	Variable Level	Number of Miss...	Percent Missing	Imputable
acct_age	INTERVAL	0	0	0
avg_arpu_3m	INTERVAL	780	1.9702	1
avg_data_chrgs_3m	INTERVAL	780	1.9702	1
avg_data_prem_chrgs_3m	INTERVAL	780	1.9702	1
avg_overage_chrgs_3m	INTERVAL	780	1.9702	1
bill_data_usg_m03	INTERVAL	0	0	0

1

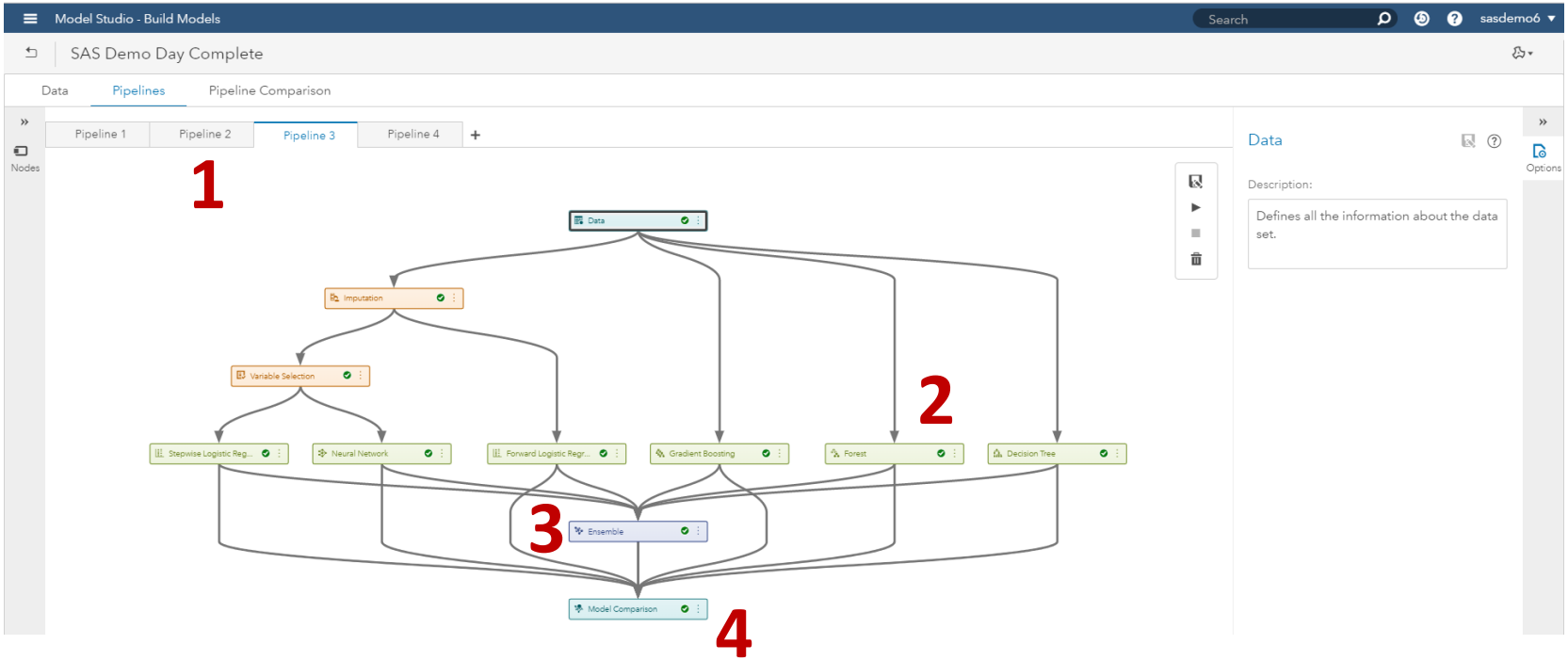
Imputed Variables Summary

Imputed Variable	Imputation Met...	Input Variable	Value	Percent Missing
IMP_mb_data_ndist_mo6m	MEAN	mb_data_ndist_mo6m	0.0455	0.2021
IMP_mou_onnet_6m_normal	MEAN	mou_onnet_6m_normal	-0.2090	37.7848
IMP_mou_onnet_pct_MOM	MEAN	mou_onnet_pct_MOM	-0.2728	47.3327
IMP_mou_total_pct_MOM	MEAN	mou_total_pct_MOM	0.0569	16.0394
IMP_tot_overage_chgs	MEAN	tot_overage_chgs	1.6132	1.9702
IMP_tot_voice_chrgs_curr	MEAN	tot_voice_chrgs_curr	34.4492	1.9702

2

Copyright © SAS Institute Inc. All rights reserved.

Больше шаблонов – богаче структура моделирования – проще итоговый выбор



Copyright © SAS Institute Inc. All rights reserved.

Спасибо за внимание!

sas.com

Copyright © SAS Institute Inc. All rights reserved.