Week 2: Prepare and Encode Data

Unit 1: Introduction to Data Preparation in SAP Predictive Analytics

SAP Run Simple

openSAP
open.sap.com

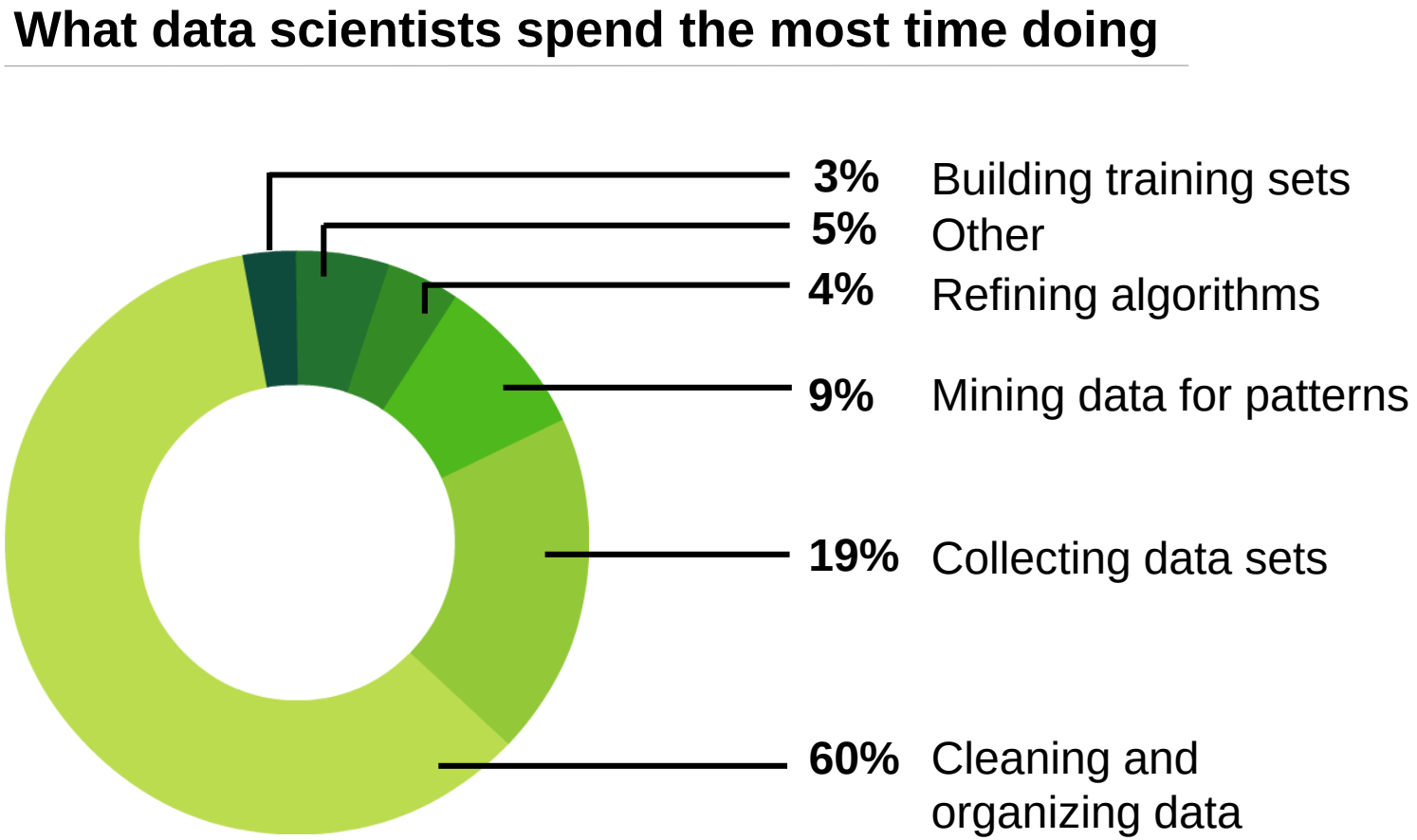
Introduction to Data Preparation in SAP Predictive Analytics

Introduction to data preparation

The chart below shows that 3 out of every 5 data scientists spend most time during their working day cleaning and organizing data.

New York Times article reported that data scientists spend from **50% to 80%** of their time mired in the more mundane task of collecting and preparing unruly digital data before it can be explored for useful nuggets.

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights. New York Times. STEVE LOHR. AUG. 17, 2014

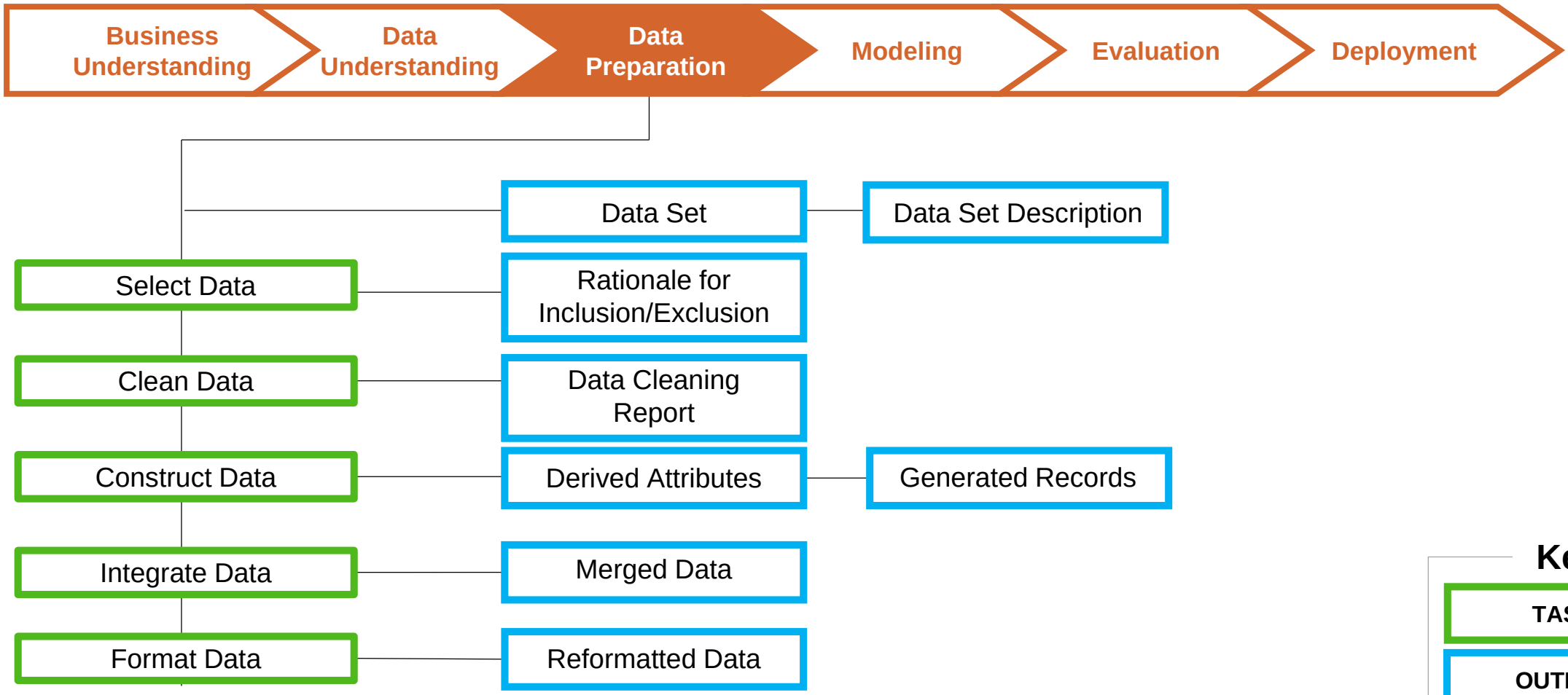


CrowdFlower Data Science Report 2016

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

CRISP-DM – Phase 3: Data Preparation



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Outputs

There are two important outputs from this phase that are not related to a specific task:

- Analytical data set
  - This is the analytical data set which will be used for modeling.
- Data set description
  - Describe the analytical data set.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Select data

- Task
  - Decide on the data to be used for analysis.
- Output – Rationale for inclusion/exclusion
  - List the data to be included/excluded and the reasons for these decisions.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Select data

- For the churn model, the telco has made the following data sources available to you:
  - A_NUMBER_FACT
  - CUSTOMER_ID_LOOKUP
  - CUSTOMER
  - CDR
  - DATA_USAGE
- All of these tables contain relevant data for the model goal: to predict churn.
- The automated modeling process in SAP Predictive Analytics will identify which attributes (i.e. the columns in the tables) are most relevant and contribute the most to the model. Any variables that do not contribute will be excluded automatically.
- Therefore, because we are using the SAP Predictive Analytics automated functionality to build the model, all of the attributes from all of the tables can be used.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Clean data

- Task
  - Raise the data quality to the level required by the selected modeling technique.
- Output – Data cleaning report
- This task is not required for this project.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Construct data

- Task
  - This task includes constructive data preparation, such as the production of derived attributes.

- Output – Derived attributes
  - Derived attributes are new attributes that are constructed from one or more existing attributes.
  - Example:

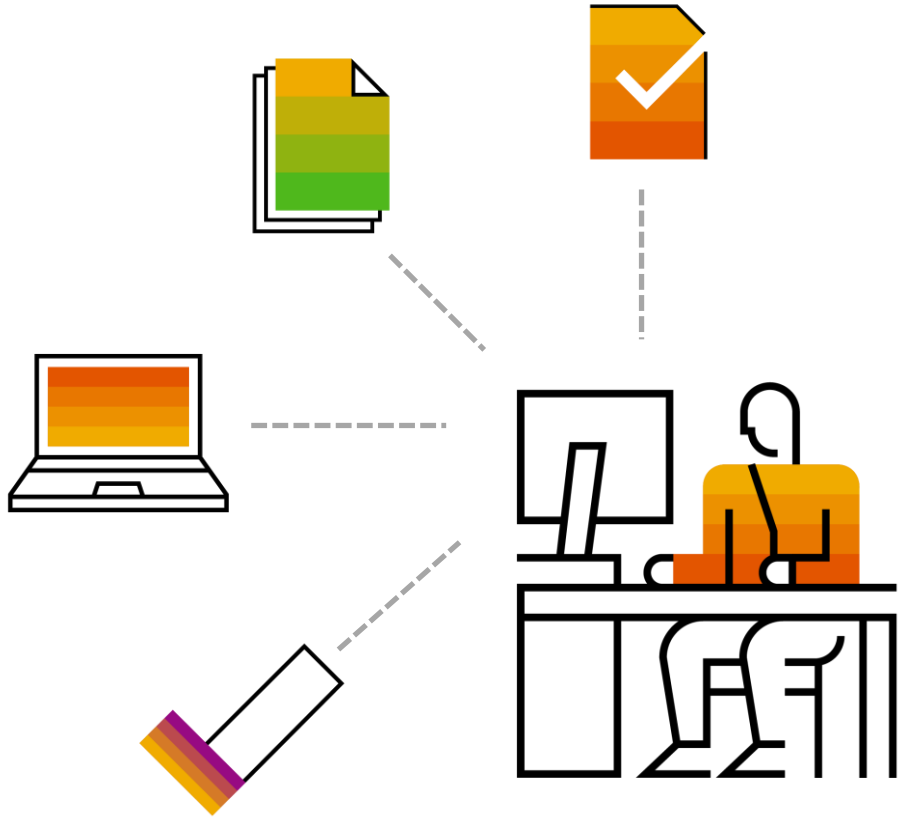
    Mean Count Voice Calls = (January Voice Calls + February Voice Calls + March Voice Calls) / 3
    Mean Duration Voice Calls = (January Voice Call Duration Total + February Voice Call Duration Total + March Voice Call Duration Total) / 3

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Integrate data

- Task
  - These are methods where information is combined from multiple tables or records
- Output – Merged data
  - “Merging” tables refers to joining together two or more tables that have different information about the same entities.
  - Merged data also covers “aggregations”.



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Merge data

	A
1	A_NUMBER
2	7809702612
3	6139214653
4	7809538328
5	7783183499
6	7788829560
7	6132919446

Table 1: A_NUMBER_FACT

	A	B
1	CUSTOMER_ID	A_NUMBER
2	1000172	2042930441
3	1000198	2502048322
4	1000210	2502164353
5	1000213	2502280241
6	1000258	2503072523
7	1000260	2503383993

Table 2: CUSTOMER_ID_LOOKUP

	A	B	C	D	E	F	G	H
1	CUSTOMER_ID	GENDER	AGE	ZIP_CODE	DISTRIBUTION_CHANNEL_ID	DEVICE_BRAND_NAME	DEVICE_MODEL_NAME	TENURE_MTHS
2	1000172	M	49	85364	XCLL001	Samsung	Galaxy S7 Edge	12
3	1000198	F	36	30032	PH00001	Apple	iPhone 7	12
4	1000210	F	24	11208	PDS0001	Samsung	Galaxy S7	12
5	1000213	F	33	10025	PDS0001	Samsung	Galaxy S7 Edge	12
6	1000258	F	55	10009	PDS0001	Huawei	Honor 8	12
7	1000260	F	36	48043	PH00001	OnePlus	3T	12
8	1000261	M	48	2155	WMN00001	Huawei	Honor 8	12

Table 3: CUSTOMER

	A	B	C	D	E	F	G	H	I
1	A_NUMBER	CUSTOMER_ID	GENDER	AGE	ZIP_CODE	DISTRIBUTION_CHANNEL_ID	DEVICE_BRAND_NAME	DEVICE_MODEL_NAME	TENURE_MTHS
2	7785585824	1000111	M	40	91706	SMO0001	OnePlus	3T	2
3	6135996086	1000112	M	62	49509	AUC0001	Google	Pixel	5
4	6136010282	1000113	F	53	11213	PDS0001	Apple	iPhone 7	5
5	6043380925	1000114	M	15	91335	WMN00001	Google	Pixel XL	10
6	7803613208	1000115	F	48	70560	SPR00001	Google	Pixel	1
7	6046152252	1000116	F	55	90650	WLM0001	Apple	iPhone 7	9
8	7803613369	1000117	M	21	90805	PDS0001	Apple	iPhone 7	1
9	6048081986	1000118	F	29	60623	SMO0001	Apple	iPhone 7	7
10	6048082683	1000119	F	57	23602	PH00001	Apple	iPhone 7	7
11	6047290312	1000120	M	32	92647	PDS0001	Apple	iPhone 7	8
12	6132266053	1000121	M	22	44107	WHP00001	Apple	iPhone 7	6

Merged Table

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Aggregating data

	A	B	C	D	E
1	A_NUMBER	B_NUMBER	TYPE	DURATION	DATE
2	7809702612	6046170793	VOICE	305	01/01/2016 03:11:00
3	6139214653	9057318773	SMS	0	01/01/2016 03:49:00
4	7809538328	7788626928	VOICE	314	01/01/2016 04:30:00
5	7783183499	6049902268	SMS	0	01/01/2016 05:45:00
6	7788829560	7053416245	SMS	0	01/01/2016 05:09:00
7	6132919446	7783894267	VOICE	137	01/01/2016 05:30:00
8	7788829560	7053416245	VOICE	105	01/01/2016 06:13:00
9	4163998288	4038305297	SMS	0	01/01/2016 06:28:00
10	7054925633	7097493780	SMS	0	01/01/2016 07:45:00
11	6047270454	6135922081	VOICE	150	01/01/2016 07:25:00
12	6134013046	7782273514	SMS	0	01/01/2016 08:41:00

CDR Table

Each VOICE, SMS and MMS call between each A_Number and B_Number has the duration and date time of the call recorded.

	A	B	C	D	E	F	G	H	I
1	A_NUMBER	TYPE_MMS_CNT_JAN	TYPE_SMS_CNT_JAN	TYPE_VOICE_CNT_JAN	CNT_JAN	TYPE_MMS_CNT_FEB	TYPE_SMS_CNT_FEB	TYPE_VOICE_CNT_FEB	CNT_FEB
2	2042930441	1	2	3	6	1	2	2	5
3	2502048322	2	3	7	12	1	2	3	6
4	2502164353	1	5	5	11	1	4	5	10
5	2503072523	5	5	17	27	3	4	12	19
6	2503383993	1	1	3	5	1	0	2	3
7	2504153759	2	2	2	6	2	2	2	6
8	2504866064	0	2	0	2	0	2	0	2
9	2505070225	1	2	2	5	0	2	2	4
10	2505162314	1	6	2	9	0	6	2	8
11	2505165087	1	0	0	1	1	0	0	1

Aggregated Table

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Format data

- Task
  - Formatting transformations refer to primarily *syntactic* modifications made to the data that do not change its meaning, but might be required by the modeling tool.
- Output – Reformatted data
  - Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.
- This task is not required for this project.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Data Preparation in SAP Predictive Analytics

Summary

- You have started the Data Preparation phase.
- All of the attributes in the tables can be used when you start to build the churn model. The SAP Predictive Analytics automated modeling process will identify which attributes are most relevant and contribute the most to the model. Any variables that do not contribute will be excluded automatically.
- You will need to build derived attributes, which are new attributes that are constructed from one or more existing attributes.
- You will also need to merge tables together, and create aggregates from the CDR table, to build the analytical data set that will be used to train and apply the churn model.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Thank you.

Contact information:

open@sap.com

© 2017 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.
See http://global.sap.com/corporate-en/legal/copyright/index.epx for additional trademark information and notices.

Week 2: Prepare and Encode Data

Unit 2: Preparing the Analytical Data Set

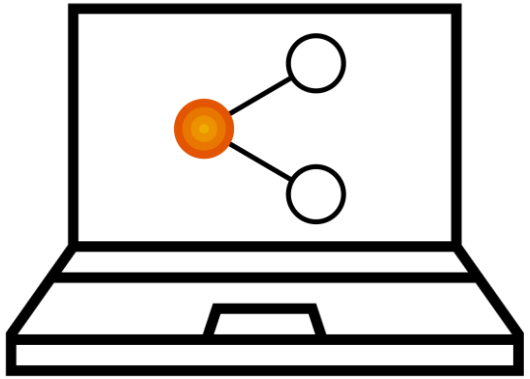
SAP Run Simple

openSAP
open.sap.com

Preparing the Analytical Data Set

Introduction

- In this unit, I'm going to show you the process to create the analytical data set (ADS) you'll use to train and apply the churn model. This data set will also be used for the cluster model, although the customer spend data will be added.
- Please follow the demo. I've also included step-by-step instructions on the following slides.



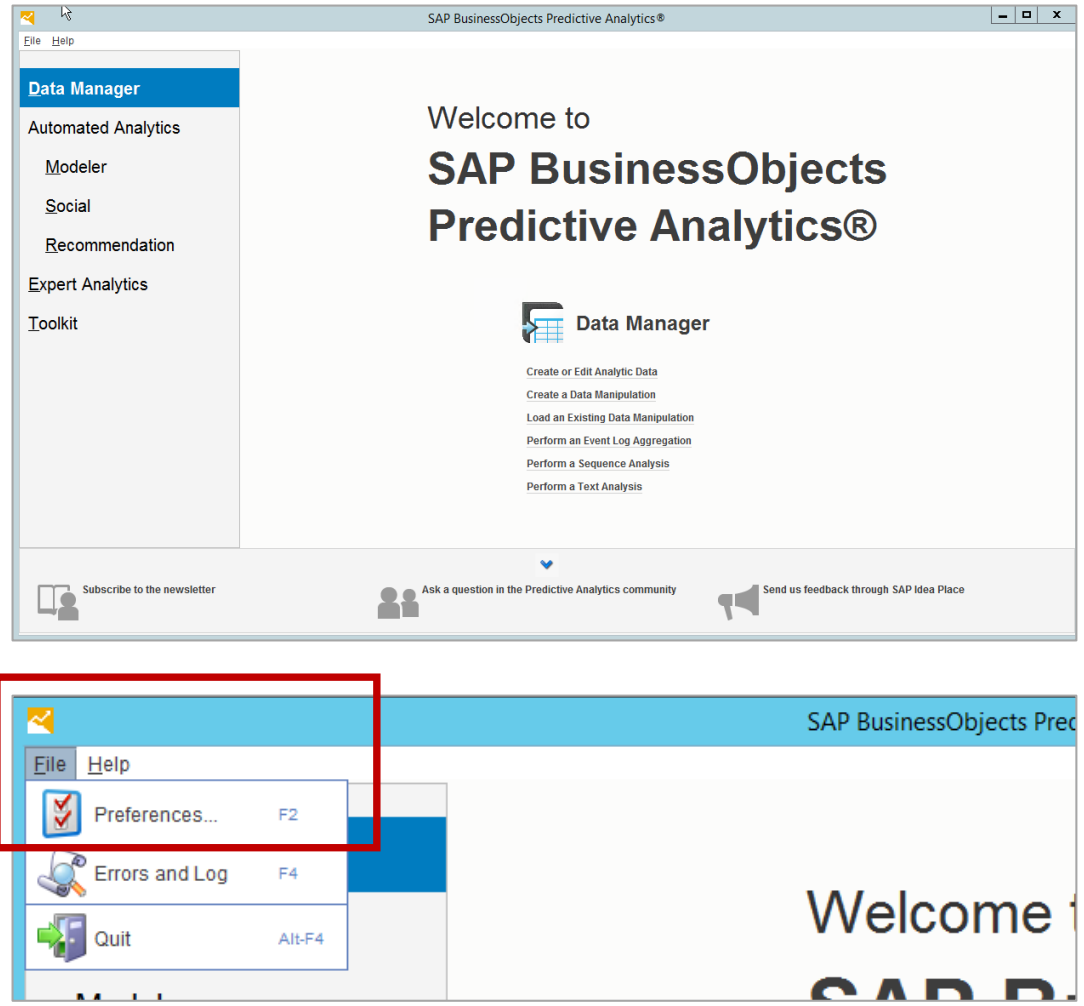
© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 1 – Create a metadata repository

- Step 1 is to create a metadata repository (MDR). This defines the location where all the metadata is stored in a single place
- Please follow these steps:

1. Access your training environment and open SAP Predictive Analytics Desktop. This will open the user interface
2. From the top left, select File \ Preferences

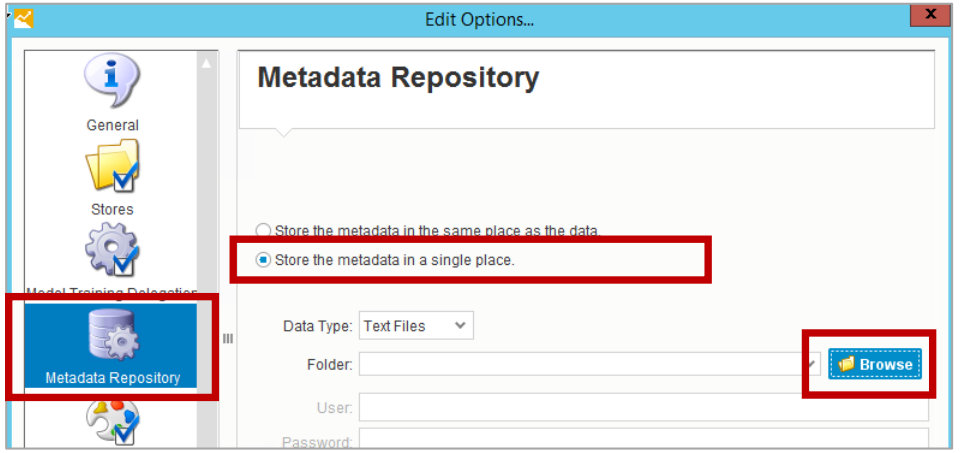


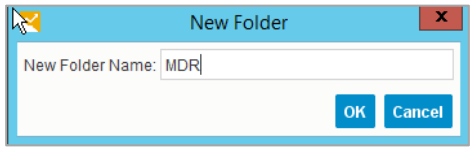
© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

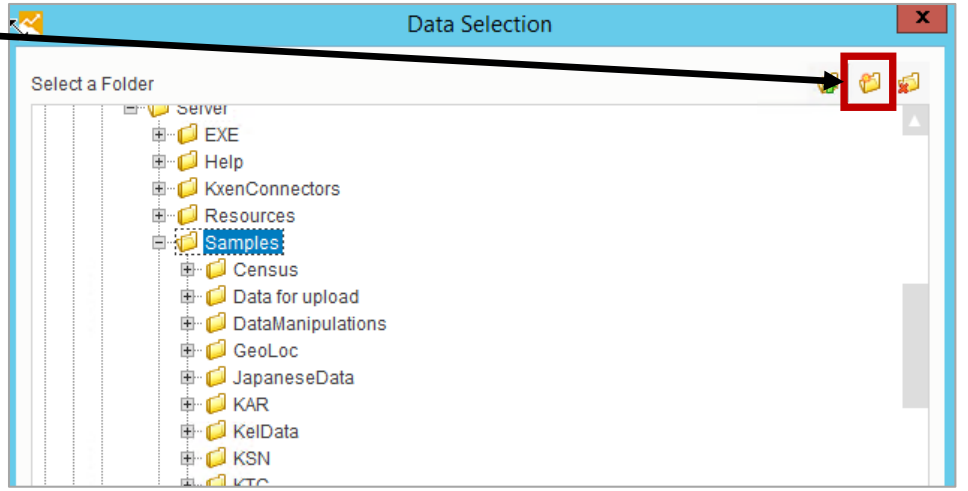
Preparing the Analytical Data Set

Step 1 – Create a metadata repository

3. In Preferences, select Metadata Repository
4. Choose “Store the metadata in a single place”
5. Under Data Type, select Text File
6. Choose Browse
7. Enter the location where the metadata repository folder will be created:
   C:\Program Files/SAP BusinessObjects Predictive Analytics/Server/Samples
8. Choose the top right radio button to “Create a new folder inside the selected one”
9. Call the new folder MDR
10. Choose OK, and the new folder will be created.
    All the metadata will now be stored here







© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 2 – Use the A_NUMBER_FACT table as the fact table and merge CUSTOMER_ID_LOOKUP table

1. Select the A_NUMBER_FACT table
   - The A_Number is a list of the unique line numbers, and is the entity for this analysis
2. Manually change A_NUMBER Type from continuous to nominal
   - A_NUMBER should be nominal, not continuous (please refer to Unit 2.5 for an explanation)
   - Note that A_NUMBER is correctly identified as a key (Key = 1)
3. Merge CUSTOMER_ID_LOOKUP table
   - Select Merge
   - Select Target Table as CUSTOMER_ID_LOOKUP
   - Join Key is A_NUMBER on the A_NUMBER_FACT table to A_NUMBER on the CUSTOMER_ID_LOOKUP table
4. View the data. Is it correct? If it is correct, save the table as ADS_1

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 3 – Merge CUSTOMER table

1. Manually change CUSTOMER_ID Type from continuous to nominal
   - CUSTOMER_ID should be nominal, not continuous (this will be explained in Unit 2.5 in more detail)
2. Merge CUSTOMER table
   - Select Merge
   - Select Target Table as CUSTOMER
   - Join Key is CUSTOMER_ID on the previously merged tables (from Step 2) to CUSTOMER_ID on the CUSTOMER table
3. ZIP_CODE should be nominal, not continuous. Manually change ZIP_CODE Type to nominal
4. TENURE_MTHS should be continuous, not nominal. Change it to continuous
5. View the data. Is it correct? If it is correct, save the table as ADS_2

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 4 – Aggregate CDR table (Voice Call Count)

1. Select New / New Aggregate
2. Select Event Table as CDR
   - The date column is automatically recognized as the DATE field in the CDR table
3. Select Join Keys
   - Reference Table Key as A_NUMBER
   - Event Table Key as A_NUMBER
4. Select the aggregation function as Count
5. Select the Target Column as * (this will count the number of calls per A_NUMBER for each month)
   - Deselect KxIndex if it has been automatically selected (note that KxIndex is a row index that is automatically added, but can be ignored in this scenario)

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 4 – Aggregate CDR table (Voice Call Count)

6. The model will be trained using 3 months of historical data, January through March 2016. The reference date for the model build is therefore midnight on the last day of March (equivalent to April 1st, 0 hours, 0 minutes, 0 seconds)
7. Go to the Period Settings tab and select Define Periods
8. Select Successive Periods
9. Select Define 3 successive period(s) of 1 Month starting 3 Month(s) before 2016-04-01 00:00:00
   - This will create aggregates for 3 months, separately, in this case for January, February and March
10. Select the Filters and Pivot Settings tab
    - Select Filter Event Table and add a filter TYPE==“VOICE”
    - We only want to create the aggregate for the voice calls, not SMS or MMS, as these are not important for this analysis

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 4 – Aggregate CDR table (Voice Call Count)

11. Choose OK
12. Save the new variable with the name AGG
13. This will create 3 new variables: AGG_3M3B_CNT_0_NoOperande, AGG_3M3B_CNT_1_NoOperande and AGG_3M3B_CNT_2_NoOperande
    - M0 refers to January, M1 refers to February, and M2 refers to March
14. Change the Alias (the name of the variable). Delete the “AGG_3M3B_” and change “NoOperande” to “VOC”, so you have CNT_0_VOC, CNT_1_VOC, CNT_2_VOC
15. View the data. Is it correct? If it is correct, save the table as ADS_3

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 5 – Aggregate CDR table (Voice Call Duration Sum)

1. Select New / New Aggregate
2. Select Event Table as CDR
   - The date column is automatically recognized as the DATE field in the CDR table
3. Select Join Keys
   - Reference Table Key as A_NUMBER
   - Event Table Key as A_NUMBER
4. Select the aggregation function as Sum (this is at the bottom of the scrolldown list of functions)
5. Select the Target Column as DURATION (this will sum the duration of the calls for each A_NUMBER for each month)
   - Deselect KxIndex if it has been automatically selected

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 5 – Aggregate CDR table (Voice Call Duration Sum)

6. The model will be trained using 3 months of historical data, January through March 2016. The reference date for the model build is therefore midnight on the last day of March (equivalent to April 1st, 0 hours, 0 minutes, 0 seconds).

7. Go to the Period Settings tab and select Define Periods

8. Select Successive Periods

9. Select Define 3 successive period(s) of 1 Month starting 3 Month(s) before 2016-04-01 00:00:00
   - This will create aggregates for 3 months, separately, in this case for January, February, and March

10. Select the Filters and Pivot Settings tab
    - Select Filter Event Table and add a filter TYPE==“VOICE”
    - We only want to create the aggregate for the voice calls, not SMS or MMS, as these are not important for this analysis

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 5 – Aggregate CDR table (Voice Call Duration Sum)

11. Choose OK

12. Save the new variable with the name AGG

13. This will create 3 new variables: AGG_3M3B_SUM_0_DURATION, AGG_3M3B_SUM_1_DURATION and AGG_3M3B_SUM_2_DURATION
    - M0 refers to January, M1 refers to February, and M2 refers to March

14. Change the Alias (the name of the variable). Delete the “AGG_3M3B_” and add suffix “VOC”, so you have SUM_0_DURATION_VOC, SUM_1_DURATION_VOC, and SUM_2_DURATION_VOC

15. View the data. Is it correct? If it is correct, save the table as ADS_4

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 6 – Derive mean count of voice calls over 3 month period

1. Select New / Expression Editor
2. Create a new variable which is the sum of the voice call count for the 3 months, divided by 3. Enter the formulae into the expression editor:
   (intToNumber(CNT_0_VOC) + intToNumber(CNT_1_VOC) + intToNumber(CNT_2_VOC)) / 3
3. Save the variable as M_MEAN_VOC_CNT, which refers to the monthly mean for voice call count
4. Use the Conversion Operators (Converts Integer to Number) to convert the integer values of the voice call count into a number. This will then create the correct output when the sum is divided by 3 to give the mean value.
5. View the data. Is it correct? If it is correct, save the table as ADS_5

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 7 – Derive mean duration of voice calls over a 3 month period

1. Select New / Expression Editor

2. Select Arithmetic Operators / Zero if NULL and create new versions for each of the three duration sum variables. The formulae is:
   ZeroIfNull(SUM_0_DURATION_VOC)
   - Save this as SUM_0_DURATION_VOC_1
   - Repeat for the other two duration sum variables
   - Make the original count variables (SUM_0_DURATION_VOC, SUM_1_DURATION_VOC, SUM_2_DURATION_VOC) invisible by deselecting these variables' Visibility

3. Create a new variable which is the sum of the voice call duration sum for the 3 months, divided by 3. Enter the formulae into the expression editor:
   (intToNumber(SUM_0_DURATION_VOC_1) + intToNumber(SUM_1_DURATION_VOC_1) + intToNumber(SUM_2_DURATION_VOC_1)) / 3

4. Save the variable as M_MEAN_VOC_DUR, which refers to the monthly mean of the voice call duration

5. View the data. Is it correct? If it is correct, save the table as ADS_6

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 8 – Derive month-on-month voice call evolutions

The change in the pattern of voice calls, month-on-month, is often a strong indicator of churn intent. One way of creating this indicator is to create a ratio that indicates the difference in call count or duration in each month, compared to the mean for the 3 months.

1. Select New / Expression Editor

2. For voice call count, derive 3 new variables:
   (CNT_0_VOC-M_MEAN_VOC_CNT)/M_MEAN_VOC_CNT and save it as CNT_0_VOC_EV
   (CNT_1_VOC-M_MEAN_VOC_CNT)/M_MEAN_VOC_CNT and save it as CNT_1_VOC_EV
   (CNT_2_VOC-M_MEAN_VOC_CNT)/M_MEAN_VOC_CNT and save it as CNT_2_VOC_EV

3. Repeat for voice call duration:
   (SUM_0_DURATION_VOC_1 - M_MEAN_VOC_DUR) / M_MEAN_VOC_DUR and save it as DUR_0_VOC_EV
   (SUM_1_DURATION_VOC_1 - M_MEAN_VOC_DUR) / M_MEAN_VOC_DUR and save it as DUR_1_VOC_EV
   (SUM_2_DURATION_VOC_1 - M_MEAN_VOC_DUR) / M_MEAN_VOC_DUR and save it as DUR_2_VOC_EV

4. View the data. Is it correct? If it is correct, save the table as ADS_7

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 9 – Merge Data_Usage table

The change in the pattern of data usage month-on-month will be a strong indicator of churn intent.

1. Merge DATA_USAGE table
   - Select Merge
   - Select Target Table as DATA_USAGE
   - Join Key is A_NUMBER on the A_NUMBER_FACT table to A_NUMBER on the DATA_USAGE table

2. Change the alias for JAN data usage to M0, FEB data usage to M1, and MAR data usage to M2. You will create the following 6 variables:

M0_Data_Usage_MB	JAN_Data_Usage_MB
M0_Data_Usage_PCT	JAN_Data_Usage_PCT
M1_Data_Usage_MB	FEB_Data_Usage_MB
M1_Data_Usage_PCT	FEB_Data_Usage_PCT
M2_Data_Usage_MB	MAR_Data_Usage_MB
M2_Data_Usage_PCT	MAR_Data_Usage_PCT

3. View the data. Is it correct? If it is correct, save the table as ADS_8

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 10 – Derive month-on-month data usage evolutions

1. Derive a new variable which is the sum of the data usage for the 3 months, divided by 3. This is the mean data usage
   - Enter the formulae into the expression editor:

     (intToNumber(M0_DATA_USAGE_MB) + intToNumber(M1_DATA_USAGE_MB) + intToNumber(M2_DATA_USAGE_MB)) / 3
   - Save the variable as M_MEAN_DATA_USAGE

2. Derive the evolution variables for each of the three months
   - Enter the formulas into the expression editor:

     (M0_DATA_USAGE_MB - M_MEAN_DATA_USAGE) / M_MEAN_DATA_USAGE and save it as DATA_0_EV

     (M1_DATA_USAGE_MB - M_MEAN_DATA_USAGE) / M_MEAN_DATA_USAGE and save it as DATA_1_EV

     (M2_DATA_USAGE_MB - M_MEAN_DATA_USAGE) / M_MEAN_DATA_USAGE and save it as DATA_2_EV

3. View the data. Is it correct? If it is correct, save the table as ADS_9

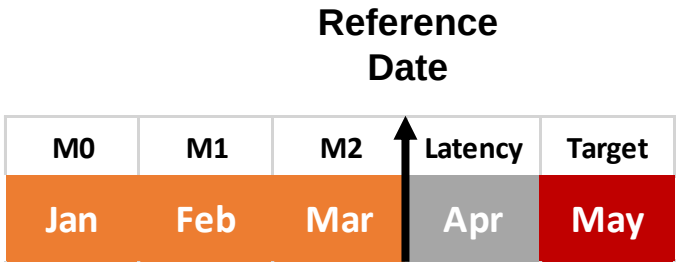
© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 11 – Create target

The DATA_USAGE table also contains a field that indicates if the customer churned in May (CHURN_MAY, 1 = churn, 0 = no churn), and similarly for June (CHURN_JUN)

1. Create a target for the model, using a condition
2. Select New / New Condition
   - Enter the following: If CHURN_MAY==1 Then 1 Else 0
   - Save this as TARGET
3. This will now provide an analytical data set where you have 3 months of history to build the model (Jan, Feb, and Mar), and a target of churn in May
   - Therefore there is a latency period of 1 month (this is the time between the end of the history and the beginning of the target period). Refer to Week 1 Unit 2 for more details
   - The Reference Date is at midnight on March 31
4. View the data. Is it correct? If it is correct, save the table as ADS_10



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 12 – Check SQL code

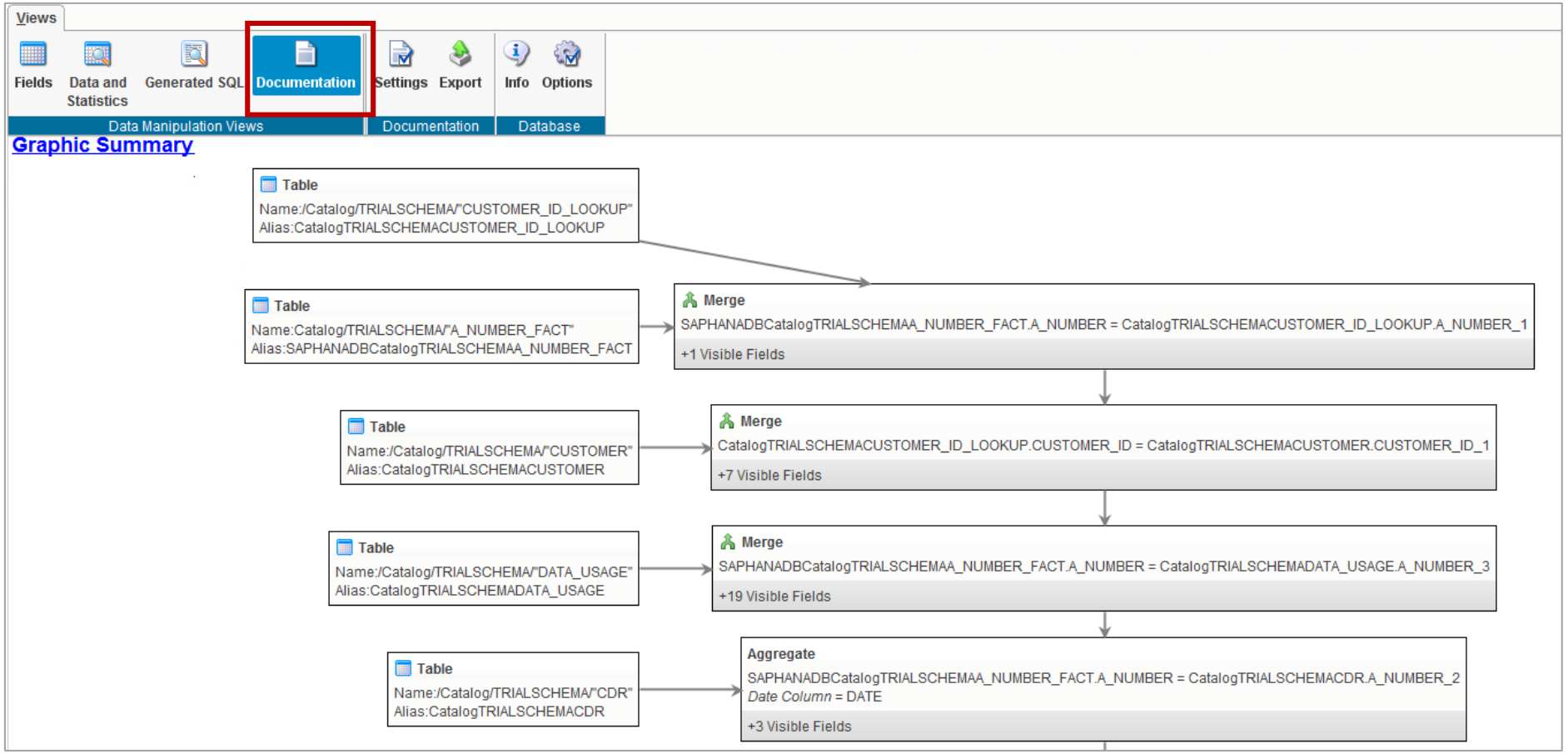
You can view the SQL code you have generated:

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Step 12 – Check the documentation

You can view the documentation that is created:



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Preparing the Analytical Data Set

Summary

- You have seen how to use the Data Manager to create a data manipulation and build an analytical data set (ADS).
- You have merged tables together, created aggregated data, and built the target for the churn model training phase.
- The Data Manager has automatically written the SAP HANA SQL code for you, and you have saved this in the metadata repository.
- The Data Manager has also created a document that explains all of the actions you have taken and details of the new variables you have created.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Thank you.

Contact information:

open@sap.com

© 2017 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.
See http://global.sap.com/corporate-en/legal/copyright/index.epx for additional trademark information and notices.

Week 2: Prepare and Encode Data

Unit 3: Introduction to Automated Modeling in SAP Predictive Analytics

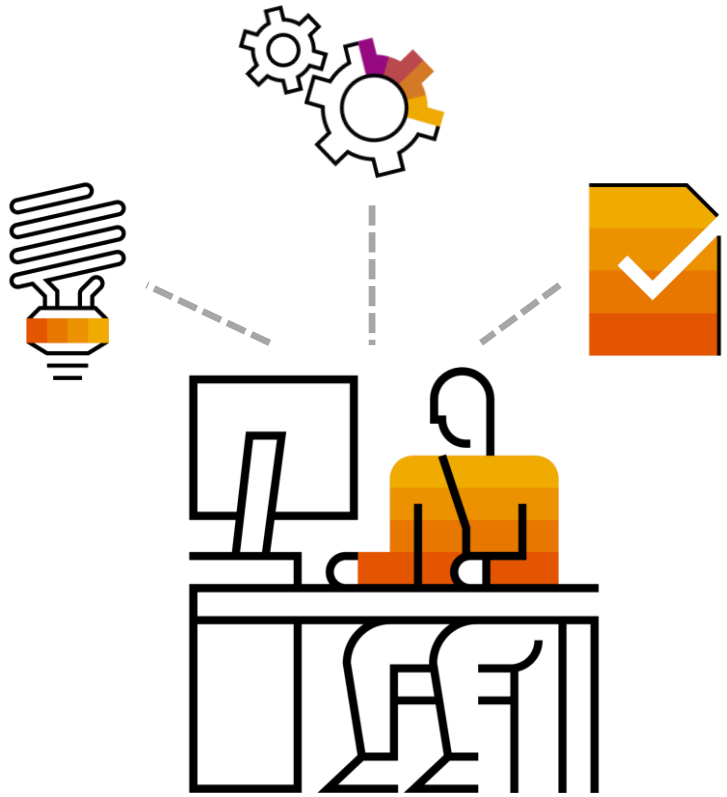
SAP Run Simple

openSAP
open.sap.com

Introduction to Automated Modeling in SAP Predictive Analytics

Introduction

- In this unit, you will learn about the benefits of using the automated modeling functionality that SAP has developed.
- You will also learn about the automated approach to predictive modeling, how it splits data to cross-validate models, avoids over-fitting models, and why you can safely use correlated explanatory variables.



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

Using predictive analytics to solve a variety of business challenges

SALES & MARKETING

- Churn reduction
- Customer acquisition
- Lead scoring
- Product recommendation
- Campaign optimization
- Customer segmentation
- Next best offer/action

OPERATIONS

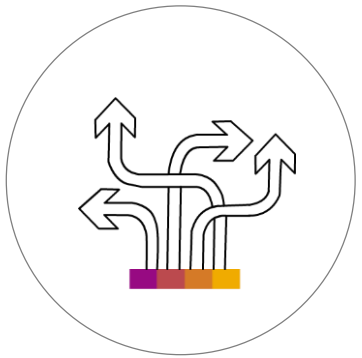
- Predictive maintenance
- Load forecasting
- Inventory/demand optimization
- Product recommendation
- Price optimization
- Manufacturing process optimization
- Quality management
- Yield management

FRAUD & RISK

- Fraud and abuse detection
- Claims analysis
- Collection and delinquency
- Credit scoring
- Operational risk modeling
- Crime threat
- Revenue and loss analysis

FINANACE & HR

- Cash flow and forecasting
- Budgeting simulation
- Profitability and margin analysis
- Financial risk modeling
- Employee retention modeling
- Succession planning



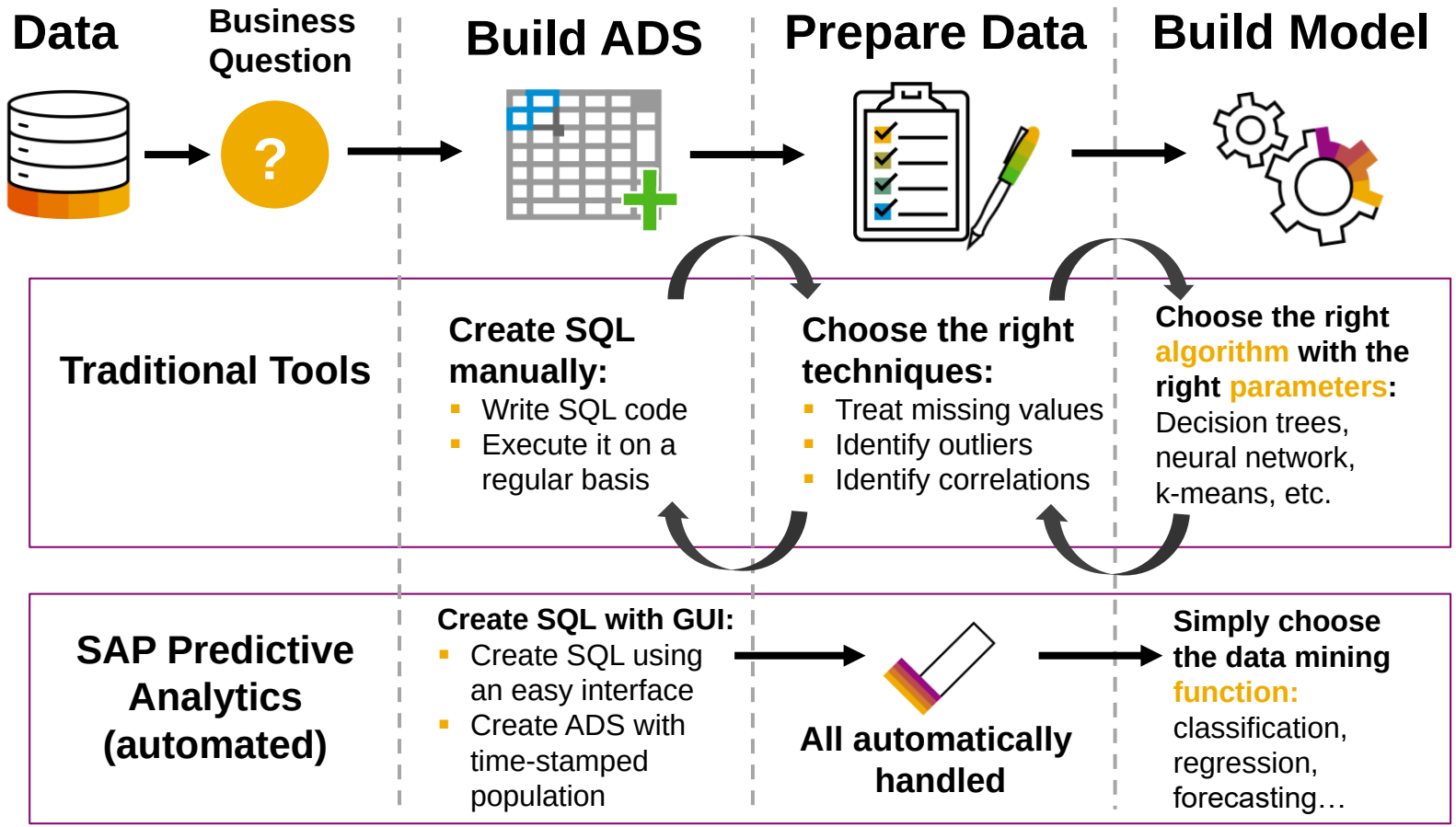
OTHER SECTORS

- Life sciences
- Healthcare
- Media
- Higher education
- Public sector / social sciences
- Construction and mining
- Travel and hospitality
- Big Data and IoT

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

Introduction – Traditional vs. automated modeling approach

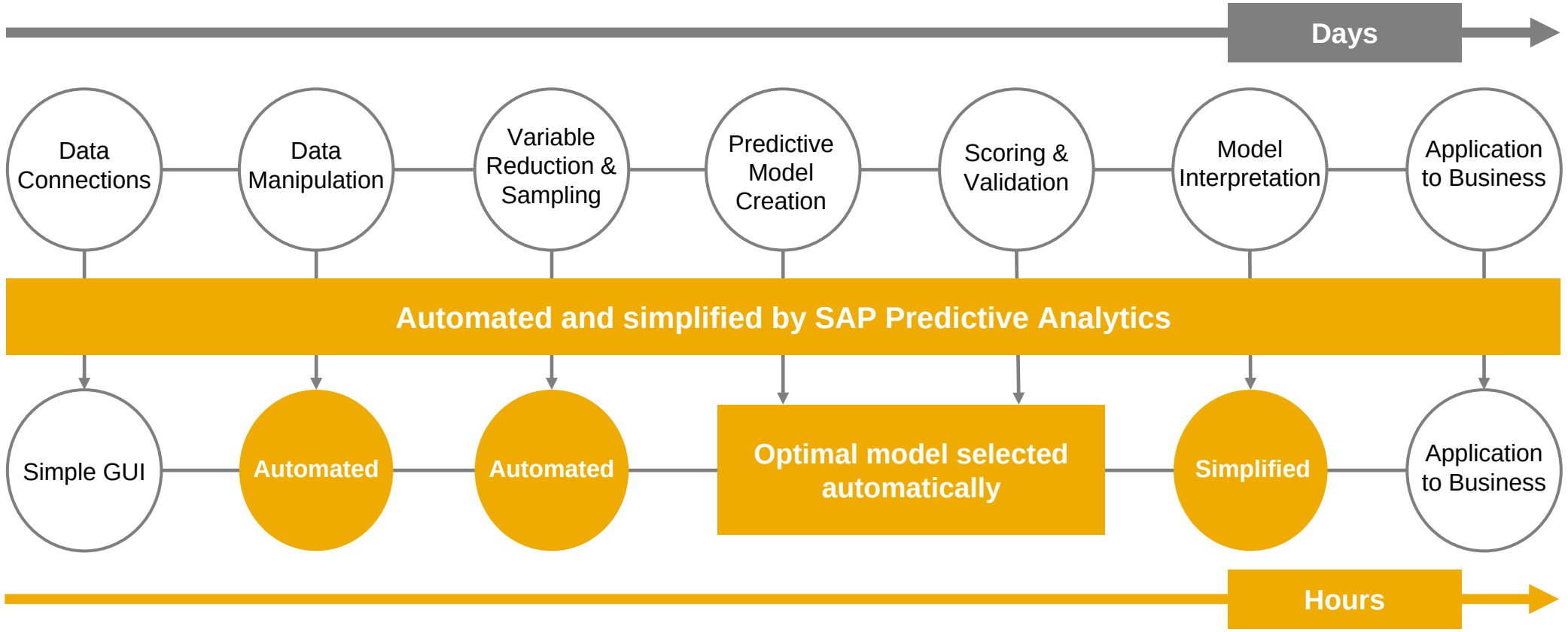


ADS = Analytical Data Set

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

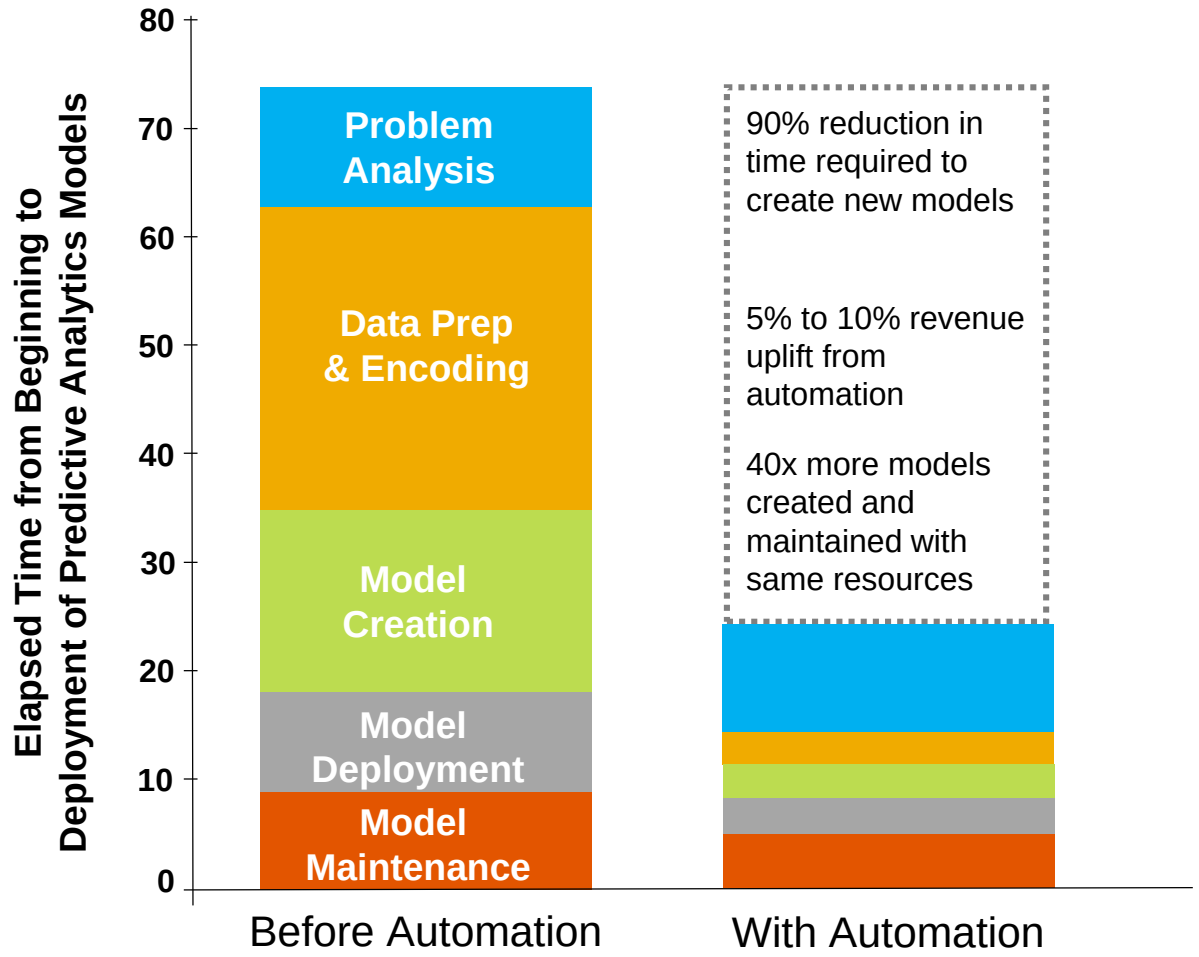
Introduction – Traditional vs. automated modeling approach



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

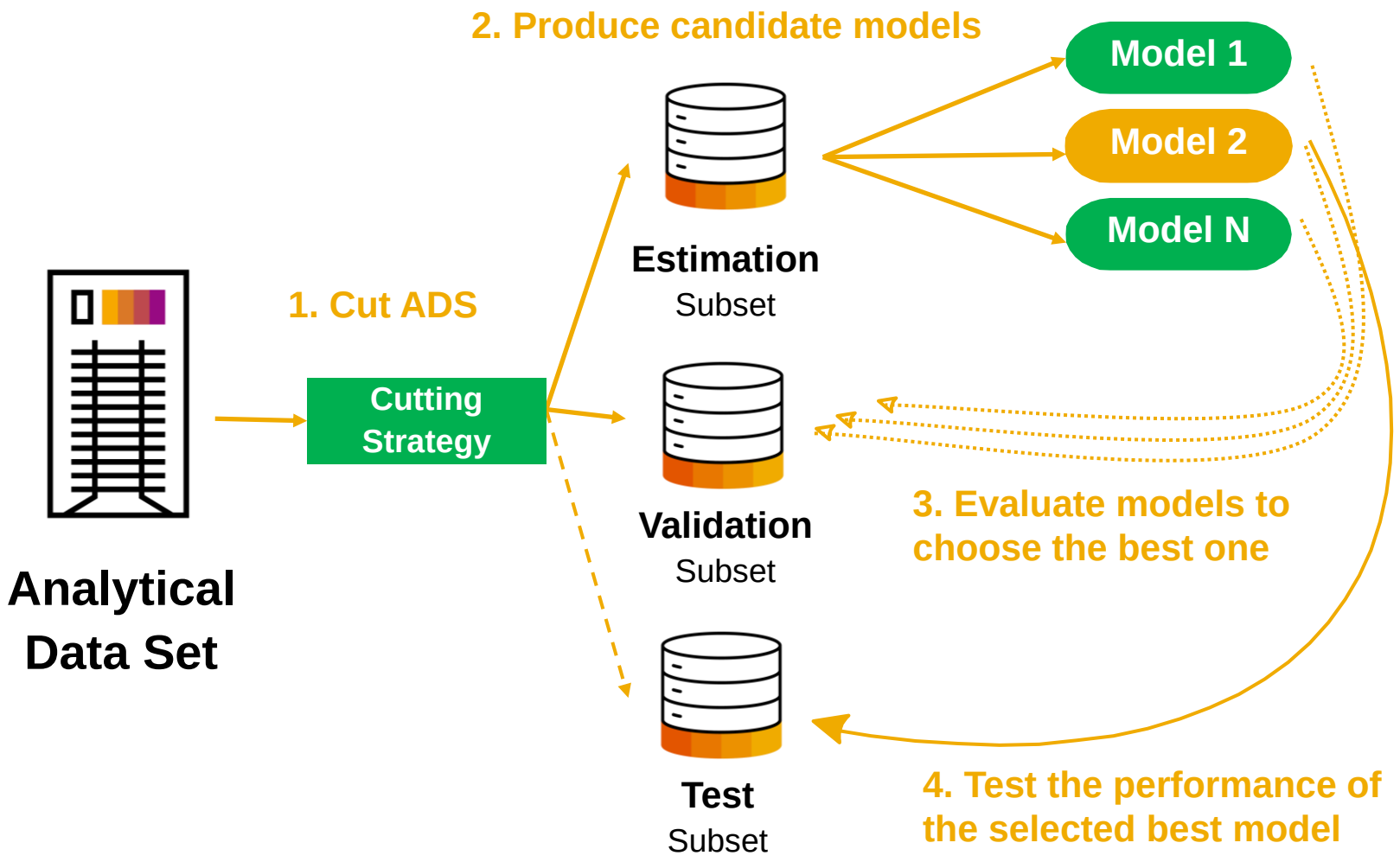
Benefits of automated modeling



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

Cutting strategy in model selection

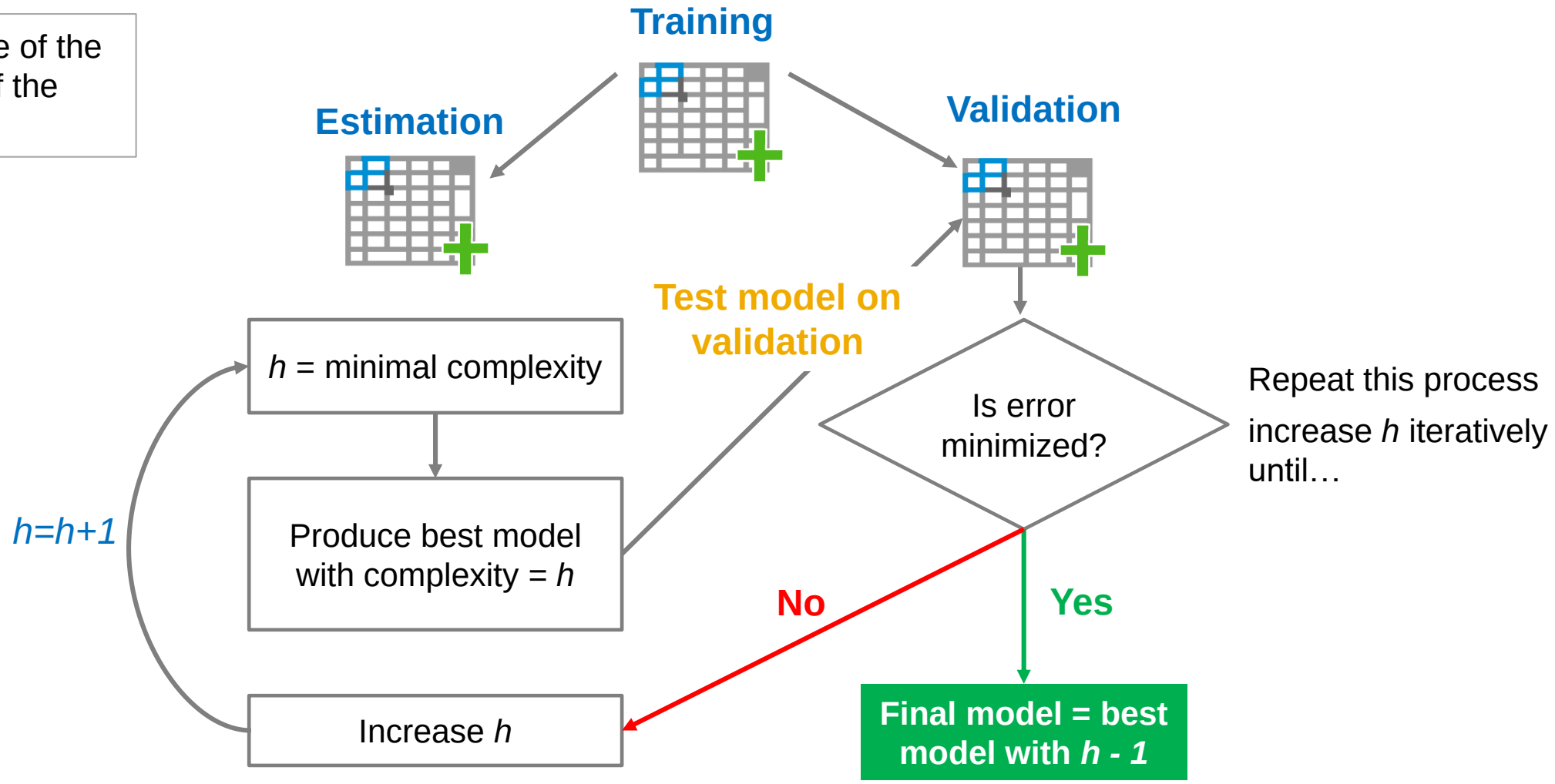


© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

Process overview

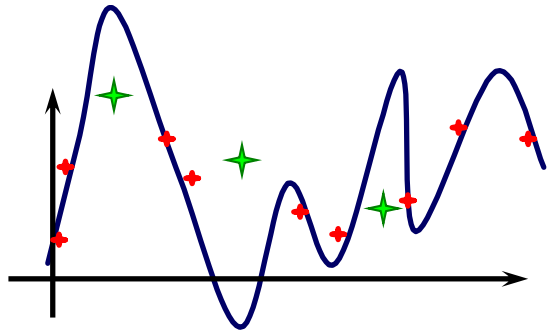
- *h* is a measure of the **complexity** of the model.



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

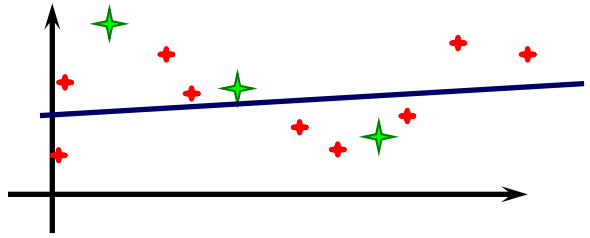
Introduction to Automated Modeling in SAP Predictive Analytics

Automatically selecting the best model



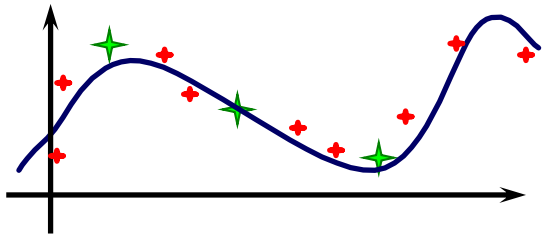
Over-Fit Model/Low Robustness

(No Training Error, High Test Error)



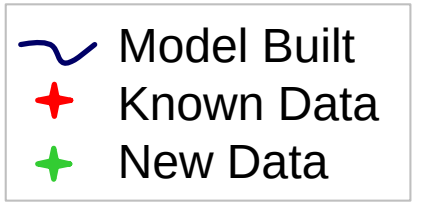
Under-Fit Model/High Robustness

(High Training Error = High Test Error)



Robust Model

(Low Training Error ≈ Low Test Error)



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

Missing values in SAP Predictive Analytics automated models

- Missing values are typically coded in data with a null value or as an empty cell, although there are a number of other representations.
- Understanding the reasons why data is missing is important to correctly handle the remaining data.
  - If values are missing completely at random, the data sample is likely to still be representative of the population.
  - But if the values are missing systematically, your analysis and models may be biased.
- Missing values in automated analytics are not excluded – they are replaced with a constant called KxMissing and then treated by the model as any other category. This allows you to assess the influence of the missing values when you have built the model.

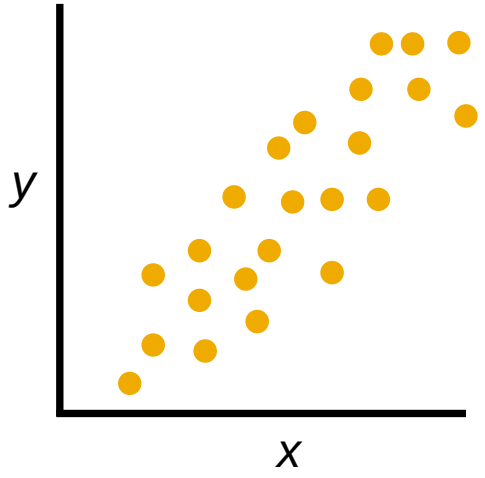
CITY
Paris
London
KxMissing
New York
KxMissing
KxMissing

https://en.wikipedia.org/wiki/Missing_data

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

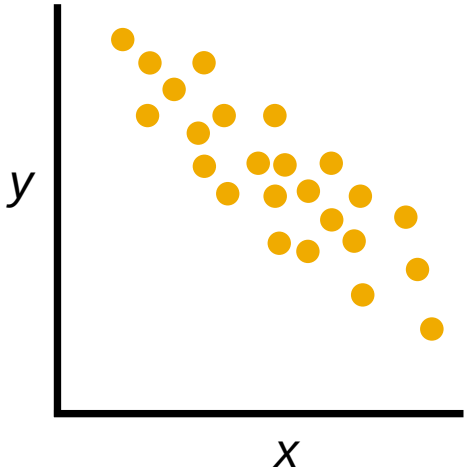
How are correlations handled in SAP Predictive Analytics automated models?



Positive correlation

The observations lie close to a straight line, which has a positive gradient.

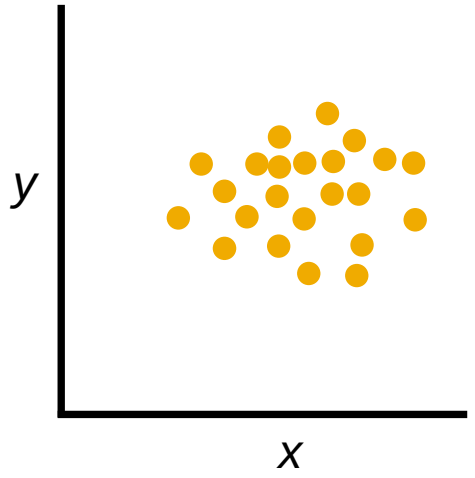
This shows that as one variable increases the other increases.



Negative correlation

The observations lie close to a straight line, which has a negative gradient.

This shows that as one variable increases the other decreases.



No correlation

There is no pattern in the observations.

This shows that there is no connection between the two variables.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Introduction to Automated Modeling in SAP Predictive Analytics

Summary

- In this unit you have learnt about:
  - The benefits of using the automated functionality in SAP Predictive Analytics
  - How the data is automatically cut into sub-samples so that the models can be cross-validated
  - How the automated tool builds and tests many models internally, so that the best model is chosen, avoiding under and over-fitting
  - How missing values do not need to be excluded from the analysis, so you can assess the influence of missing data when you have built the model
  - How the automated modeler uses a ridge regression approach, so you don’t need to de-correlate the explanatory variables

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Thank you.

Contact information:

open@sap.com

© 2017 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.
See http://global.sap.com/corporate-en/legal/copyright/index.epx for additional trademark information and notices.

Week 2: Prepare and Encode Data

Unit 4: Automated Data Encoding

SAP Run Simple

openSAP
open.sap.com

Automated Data Encoding

Introduction

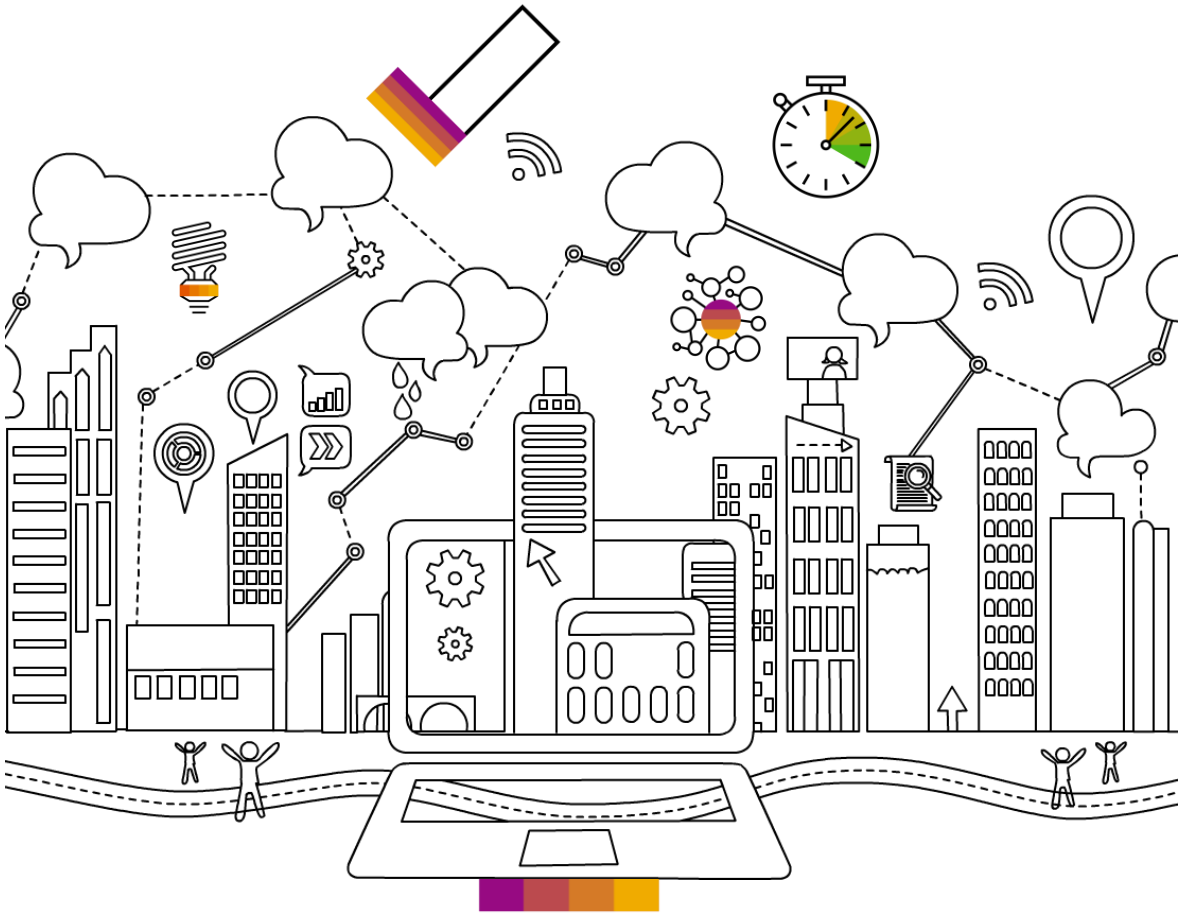
- Data encoding is an essential part of the data preparation process.
- The data encoding process prepares missing values in the data, deals with outliers, and creates data bins or bands to transform raw data into a “mineable” source of information.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Automated Data Encoding

Benefits of using an automated approach

- With traditional approaches, the process of encoding each explanatory variable prior to model development accounts for a large portion of the analysis time.
- The data encoding component in automated SAP Predictive Analytics quickly and automatically transforms raw data into a “mineable” source of information.
- It uses a “cross-validation” approach.
- As well as this automated approach, manual encoding is also available if required.



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Automated Data Encoding

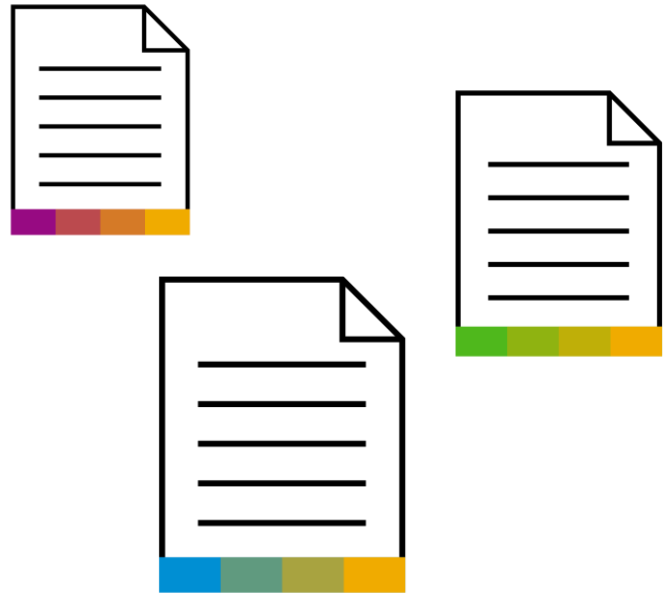
Different variable types

There are three different types of variable:

1. **Nominal variables**
   – A discrete and unordered set of values or categories

2. **Ordinal variables**
   – A discrete and ordered set of values

3. **Continuous variables**
   – A real number that can have any value (with fractions/decimal places)



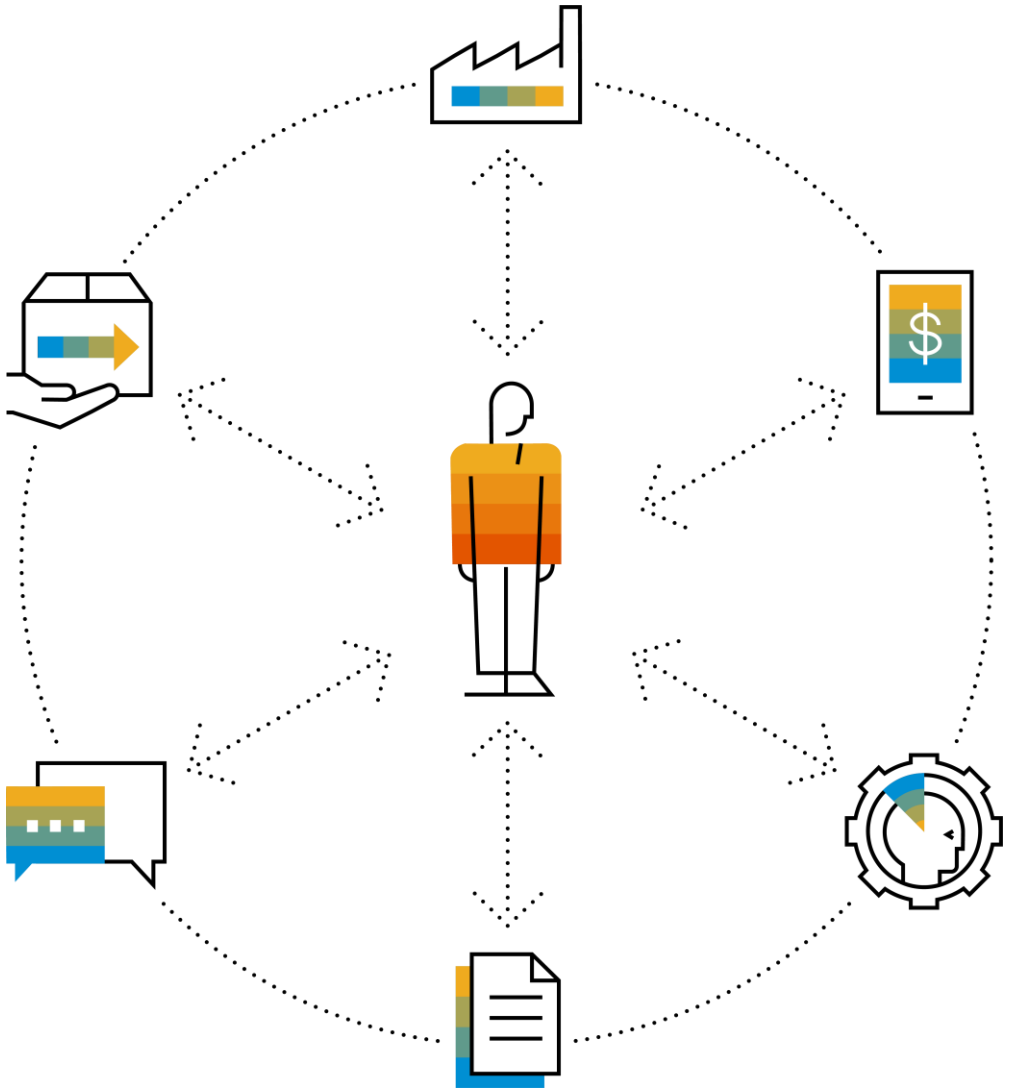
© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Automated Data Encoding

Nominal variable

- A nominal variable is a discrete (categorical), qualitative variable that characterizes, describes, or names an element of a population.
- Examples:
  - Hair color (brown, blond, ginger…)
  - Make of car (Mercedes, Ford...)
  - Gender (male, female)
  - Postal (ZIP) code
  - Residence city (London, New York, Paris…)

The order of the categories **does not** matter



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

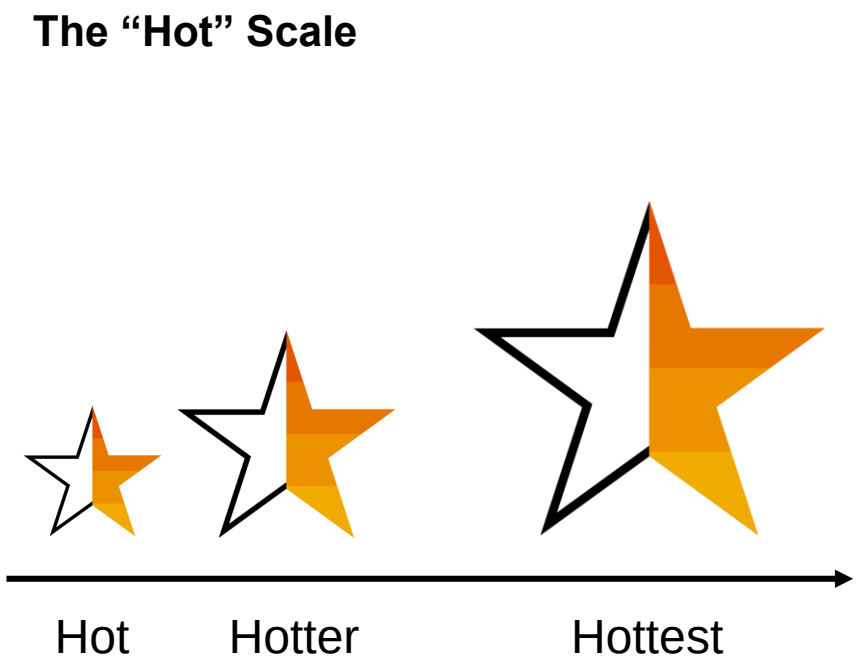
Automated Data Encoding

Ordinal variable

- An ordinal variable is a discrete (categorical), qualitative variable that has order.

- Examples:
  - Gold, silver, bronze
  - Satisfaction level (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied)
  - Pain level (mild, moderate, severe)

Note: The order of the categories **does** matter



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

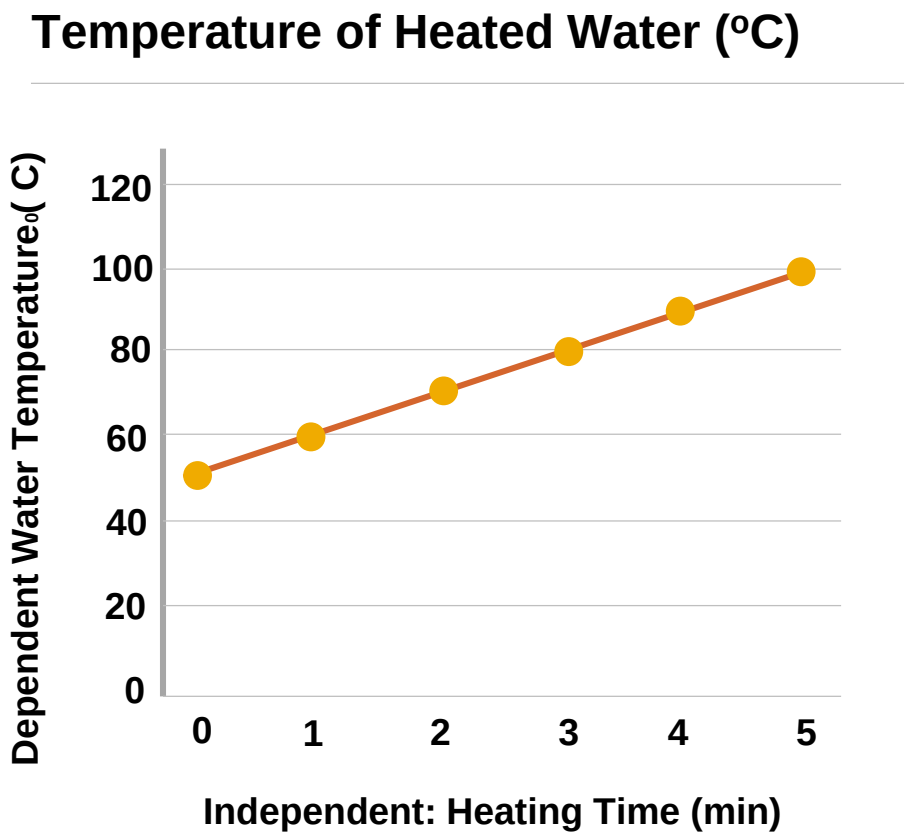
Automated Data Encoding

Continuous variable

- A continuous variable is a quantitative variable.
- It is a real number that can take any value (with fractions/ decimal places) between two specific numbers.
- It accommodates all basic arithmetic operations (addition, subtraction, multiplication, and division).

Examples:

- Income ($)
- Age (years)
- Running time (minutes)
- Bank account balance ($)
- Distance (miles)
- Any ratio or calculated value

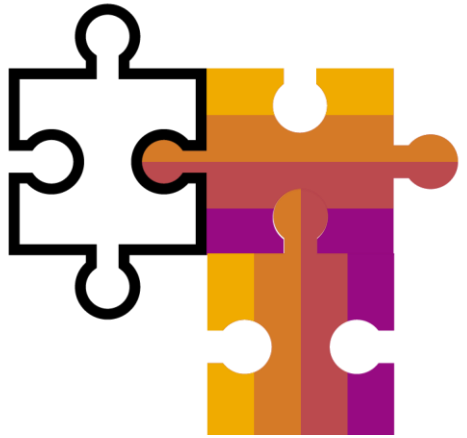


© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Automated Data Encoding

Missing values

- A missing value is an empty cell in your data set.
- Missing values in a data set can be due to error or because they are simply not available.
- They can be removed from the data set, estimated, or kept.
- The analysis could also be stopped so that further investigation of the reason for missing values can be undertaken.

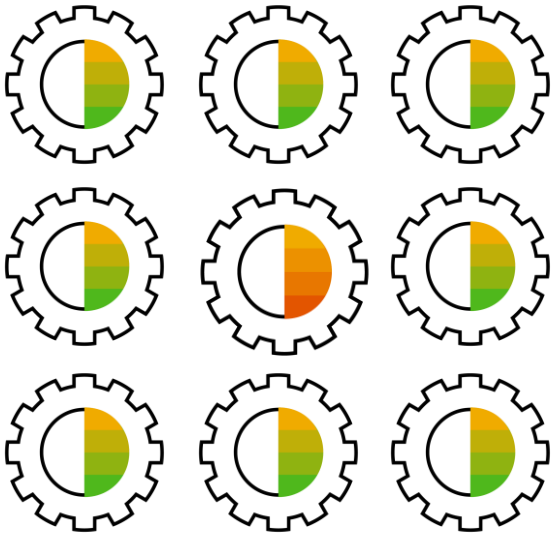


© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Automated Data Encoding

Outliers

- For a continuous variable – An outlier is a single or low-frequency occurrence of the value of a variable that is far from the mean as well as the majority of other values for that variable.
- For a categorical variable (nominal or ordinal) – An outlier is a single or very low-frequency occurrence of a category of a variable.



© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Automated Data Encoding

Summary

- In this unit, you have learnt about the automated encoding functionality that uses a cross-validation approach to transform raw data into “minable” data.
- It is important that you understand the difference between nominal, ordinal, and continuous variables, as different encoding algorithms are applied to each type.
- Missing values and outliers are automatically dealt with when you use the automated encoding functionality.

© 2017 SAP SE or an SAP affiliate company. All rights reserved. I PUBLIC

Thank you.

Contact information:

open@sap.com

© 2017 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.
See http://global.sap.com/corporate-en/legal/copyright/index.epx for additional trademark information and notices.