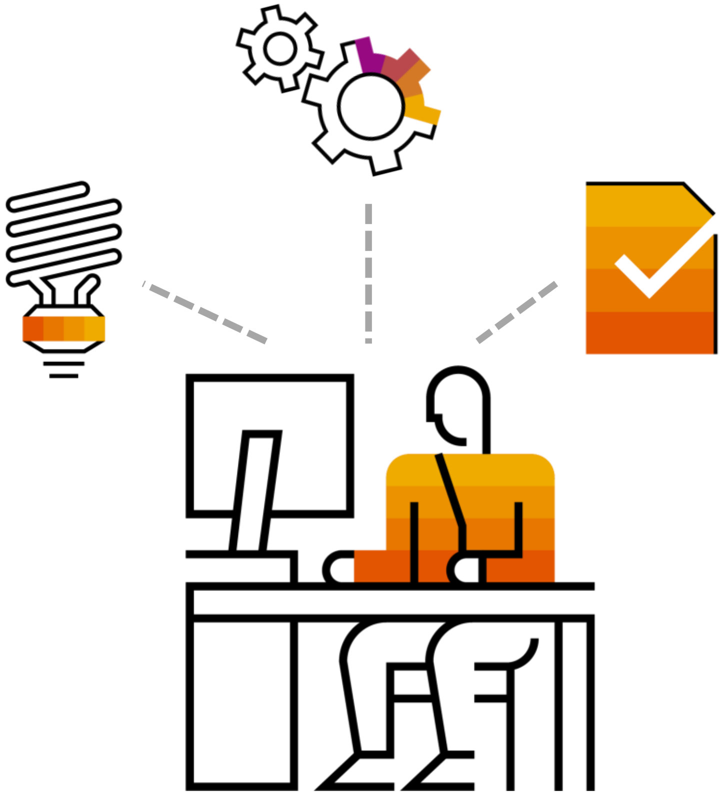
Week 1: Case Study Introduction

Unit 1: CRISP-DM Project Methodology – Recap

CRISP-DM Project Methodology – Recap

Data Science in Action – The next 4 weeks

What to expect in the next 4 weeks



CRISP-DM Project Methodology – Recap

Curriculum flow (weeks 1-2)

1

Case Study Introduction

- CRISP-DM Project Methodology – Recap
- Introduction to the Telco Case Study
- Understanding the Business Requirements
- Understanding the Data

Weekly Assignment

2

Prepare and Encode Data

- Introduction to Data Preparation in SAP Predictive Analytics
- Preparing the Analytical Data Set
- Introduction to Automated Modeling in SAP Predictive Analytics
- Initial Data Analysis
- Automated Data Encoding

Weekly Assignment

CRISP-DM Project Methodology – Recap

Curriculum flow (weeks 3-4)

3

Develop, Evaluate, and Deploy Models

- Data Description and Data Roles
- Developing an Initial Churn Model
- Evaluating the Initial Churn Model
- Deploying the Initial Model Using SAP Predictive Analytics

Weekly Assignment

4

Monitor Models and Improve Performance

- Monitoring and Maintaining Performance with Predictive Factory
- Improving the Model – Developing a Social Link Analysis
- Introduction to Segmentation
- Developing a Segmentation Using SAP Predictive Analytics
- Wrap-Up

Weekly Assignment

Final Exam

CRISP-DM Project Methodology – Recap

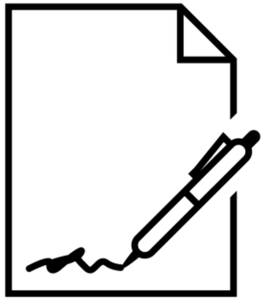
Cumulative points lead to record of achievement

Watch the deadlines! !

Participate in Weekly Assignment (Weeks1-4)

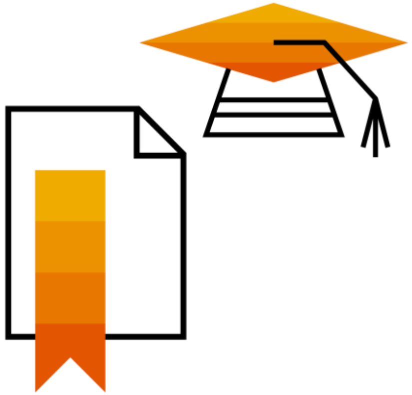
4 assignments
4 x 30 = 120 points

Final Exam
(Week 5)



120 points

Record of Achievement

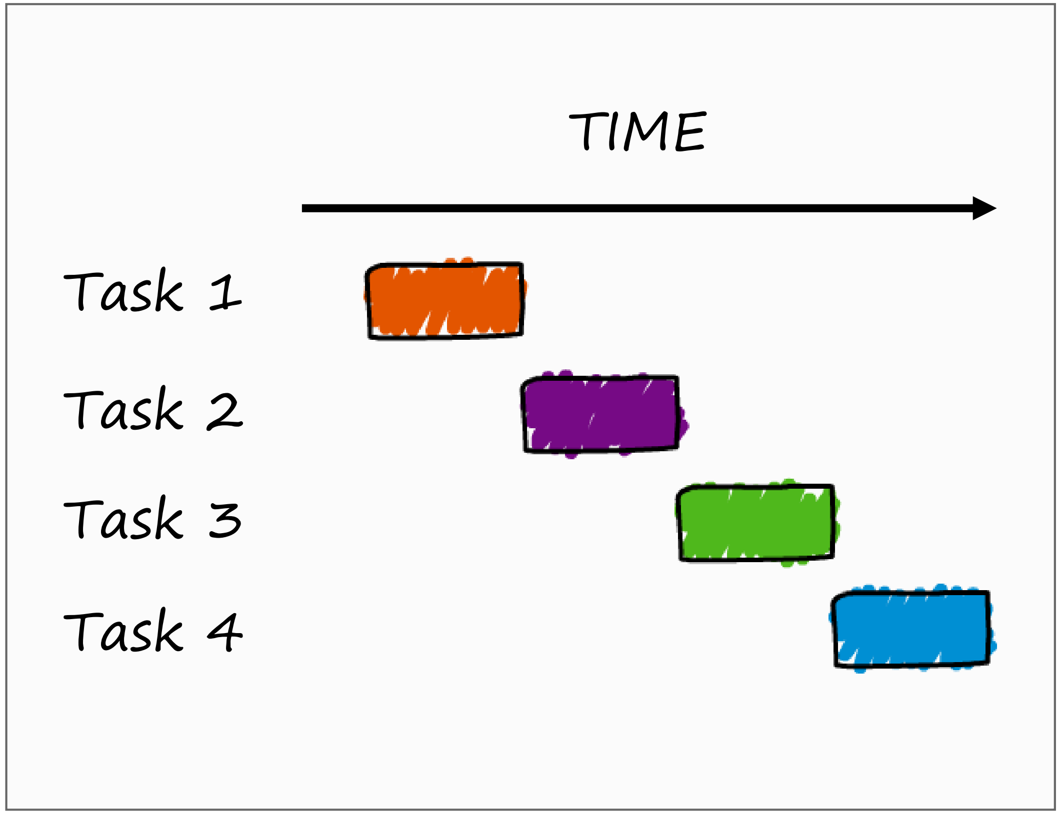


When results above
120 points

CRISP-DM Project Methodology – Recap

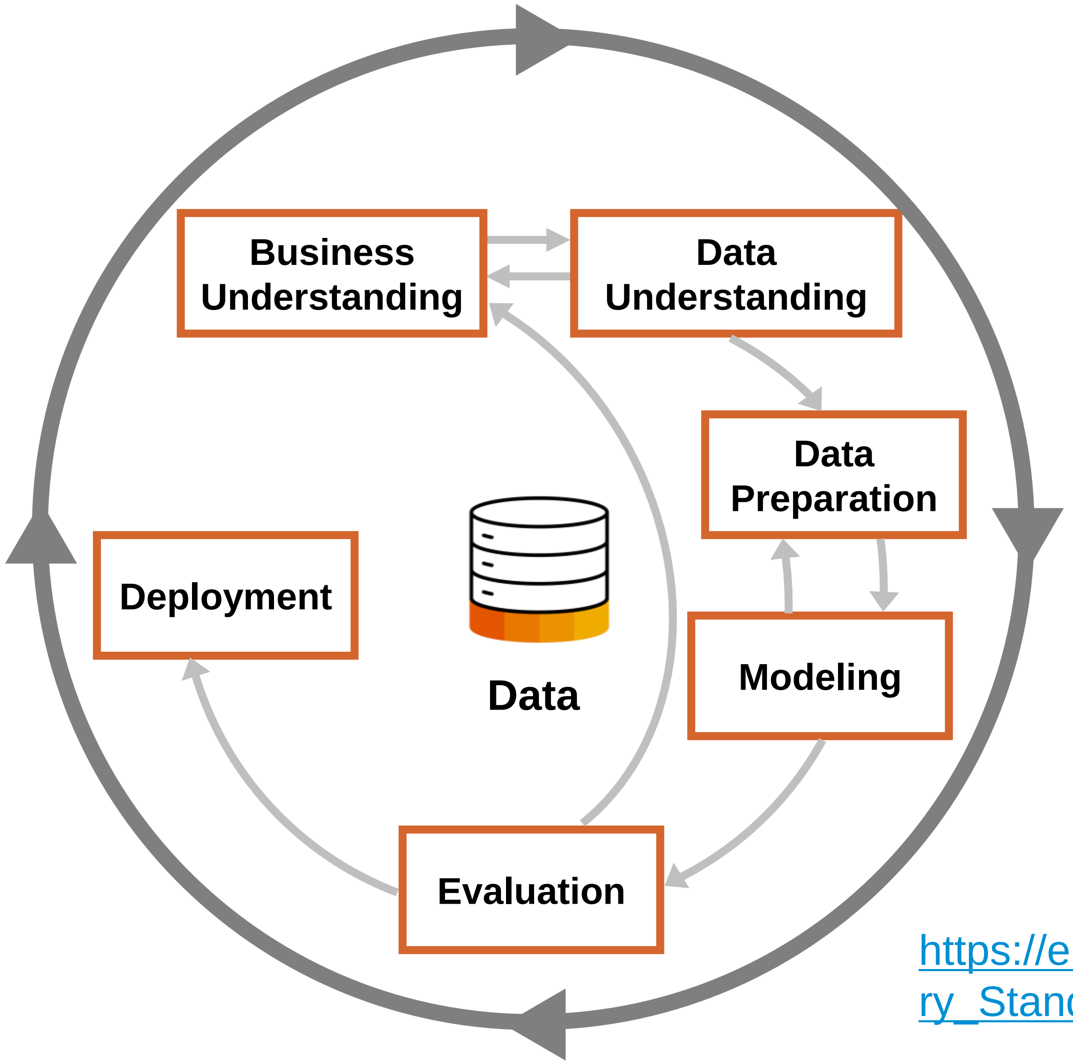
Why should there be a project methodology?

- It is important to have a clearly defined process of initiating, planning, executing, controlling, and closing the work of a data science team to achieve the specific project goals and meet the specific success criteria.
- A project methodology:
  - Provides a clear process framework so that project goals and success criteria an be achieved
  - Allows projects to be replicated
  - Provides an aid to project planning and management
  - Is a “comfort factor” for new adopters



CRISP-DM Project Methodology – Recap

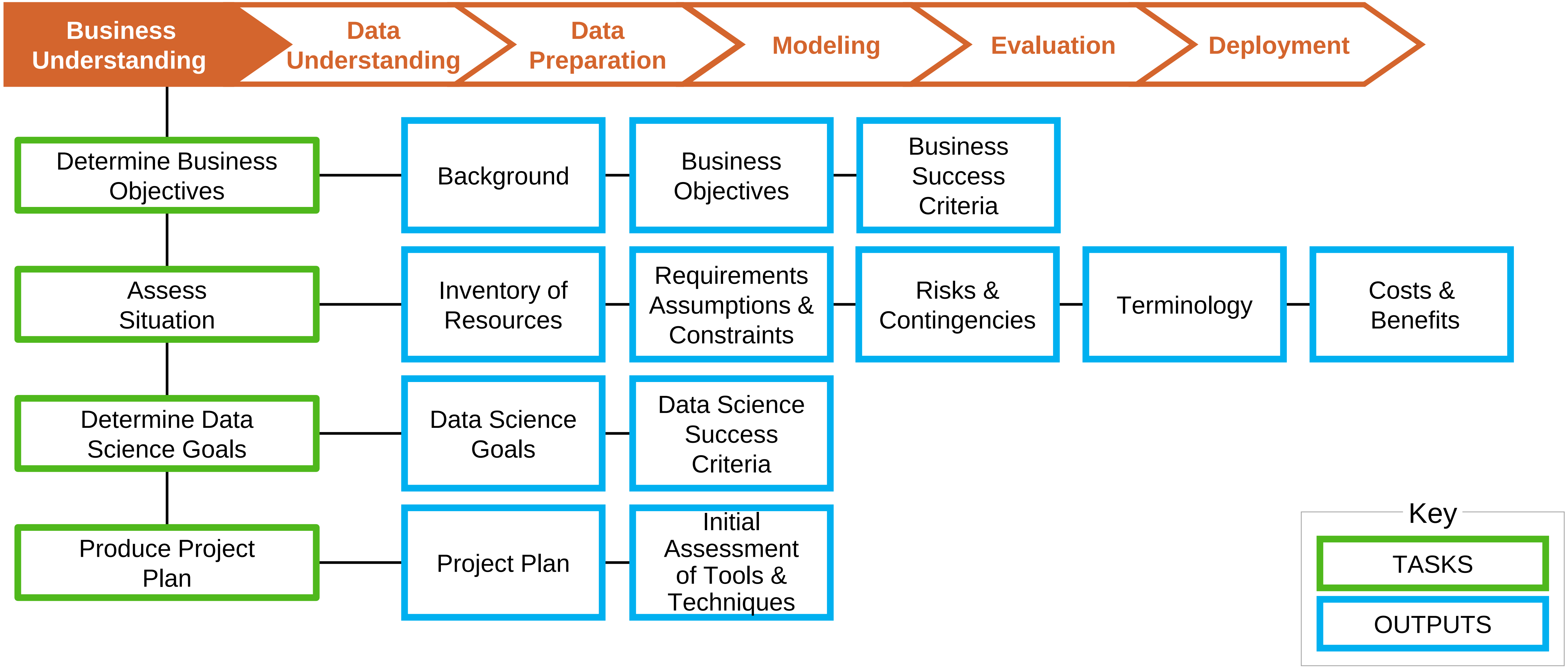
Cross-industry standard process for data mining (CRISP-DM)



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

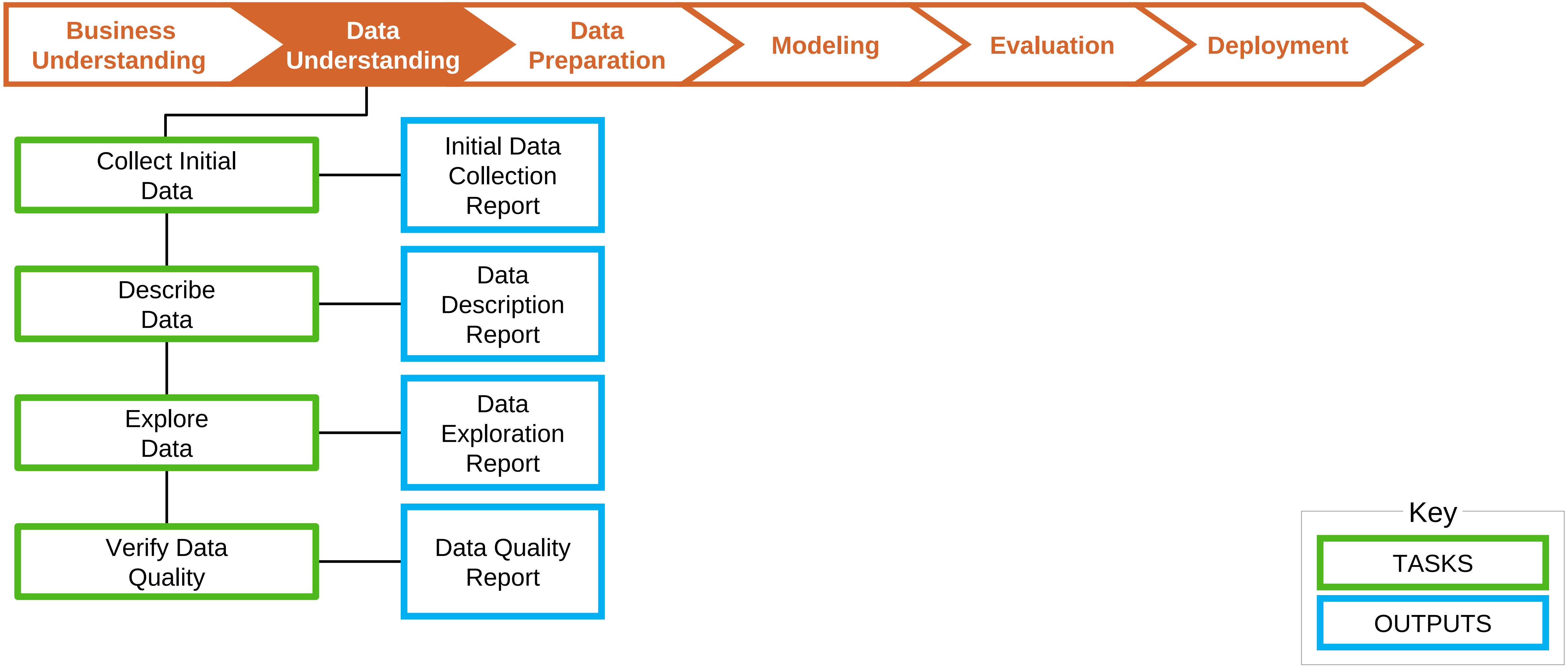
CRISP-DM Project Methodology – Recap

CRISP-DM – Phase 1: Business Understanding



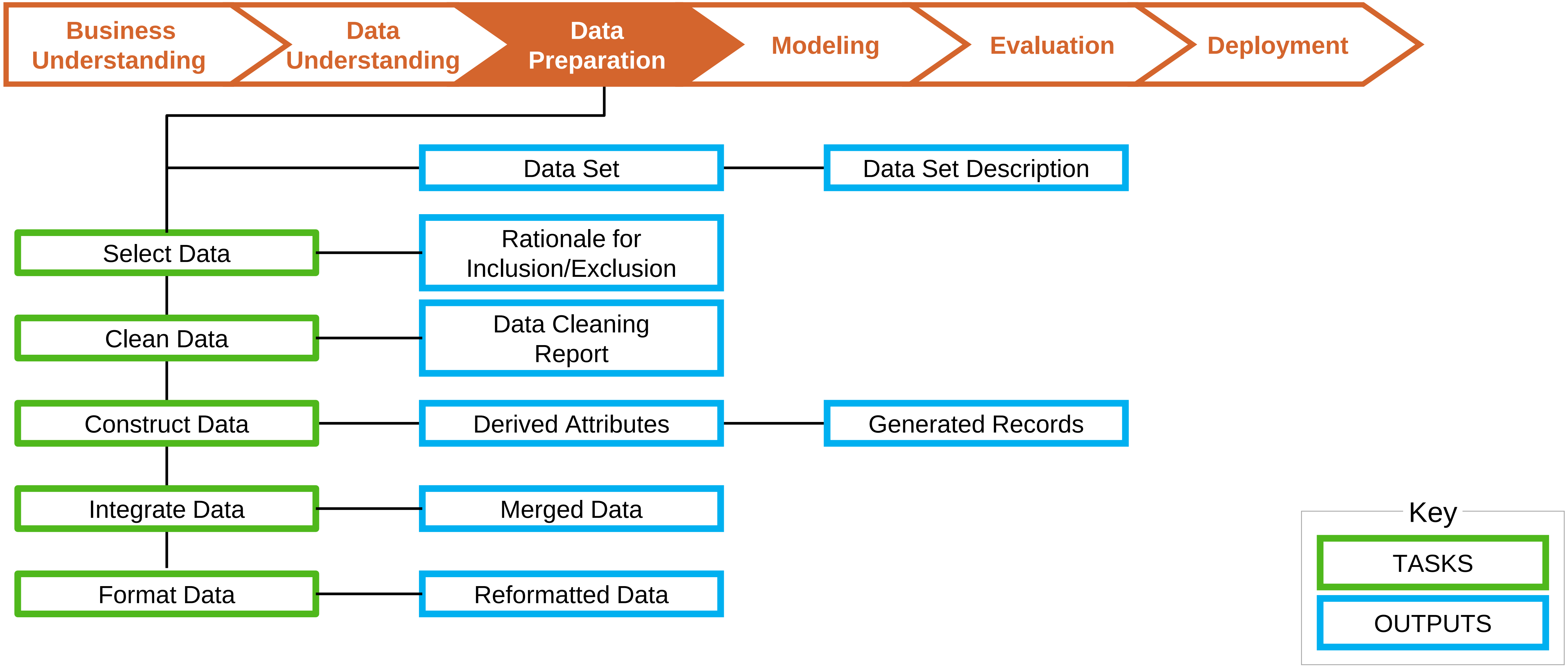
CRISP-DM Project Methodology – Recap

CRISP-DM – Phase 2: Data Understanding



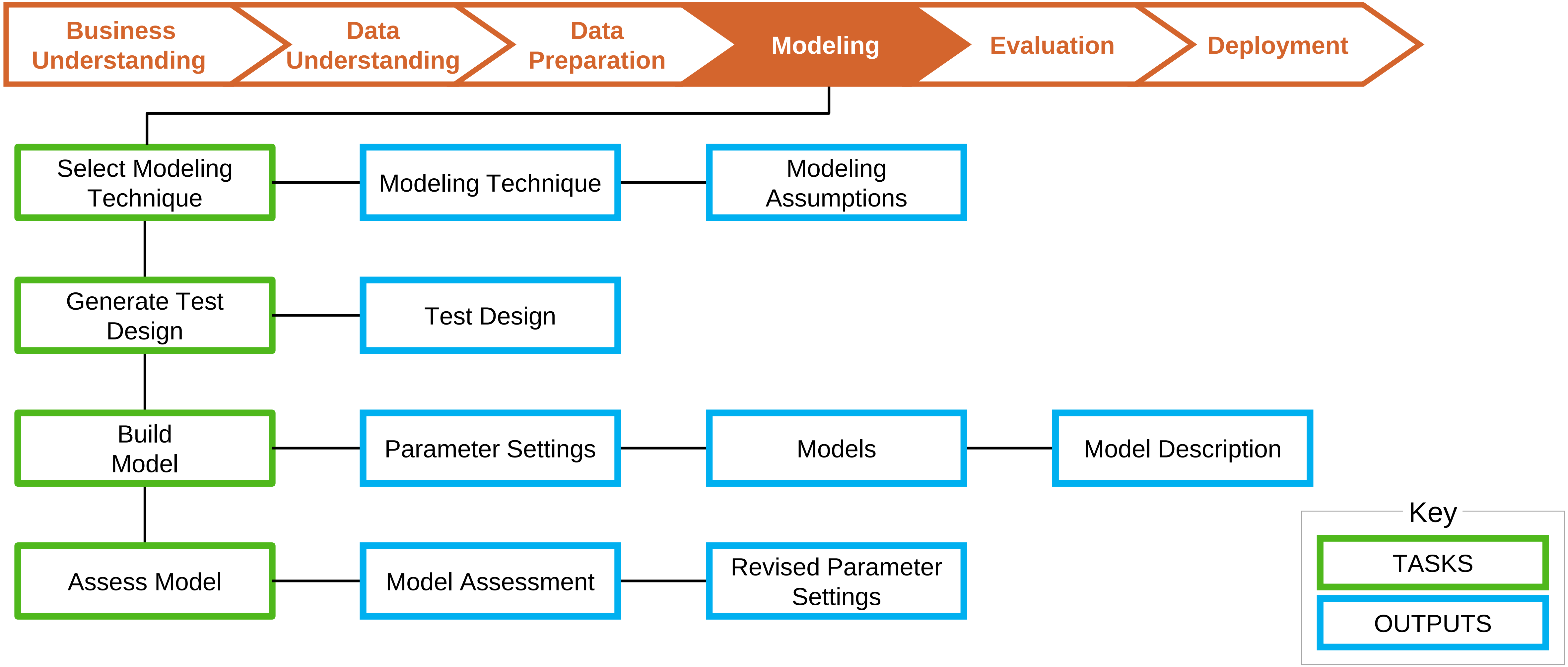
CRISP-DM Project Methodology – Recap

CRISP-DM – Phase 3: Data Preparation



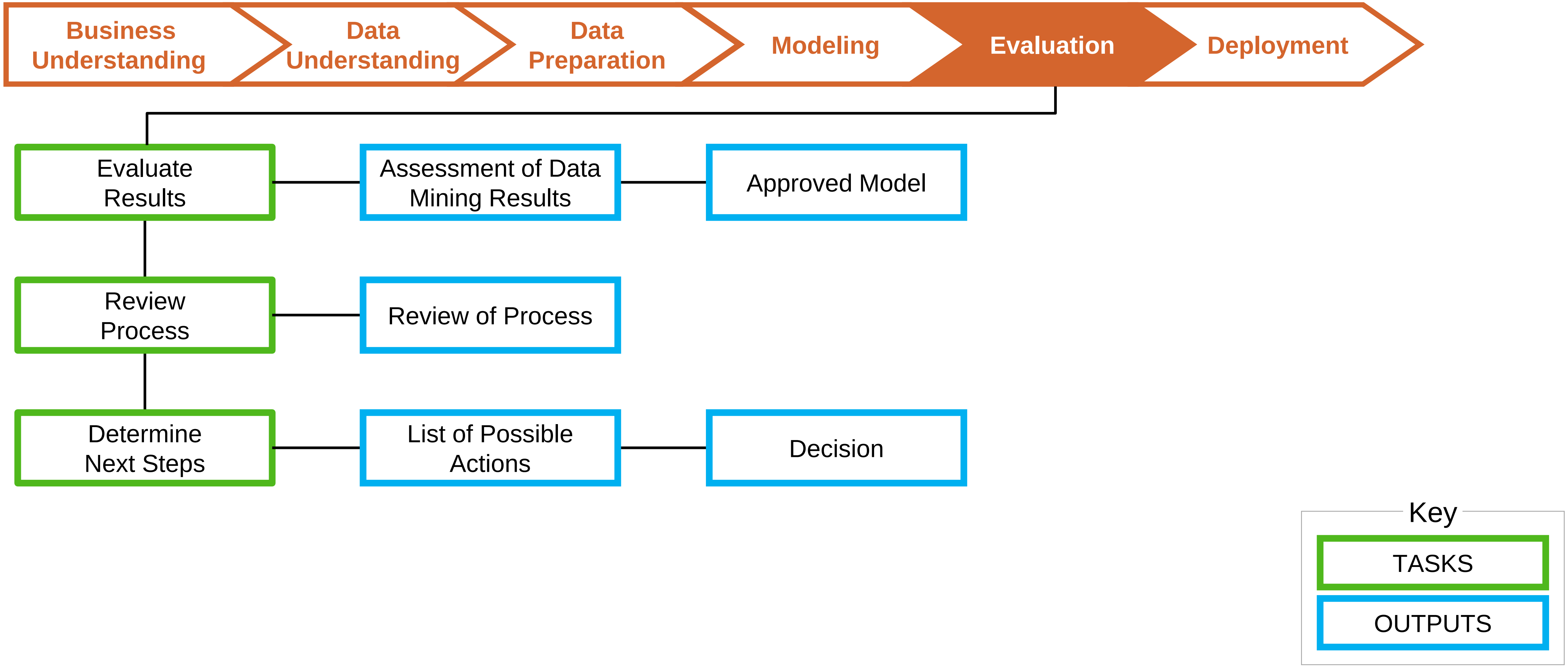
CRISP-DM Project Methodology – Recap

CRISP-DM – Phase 4: Modeling



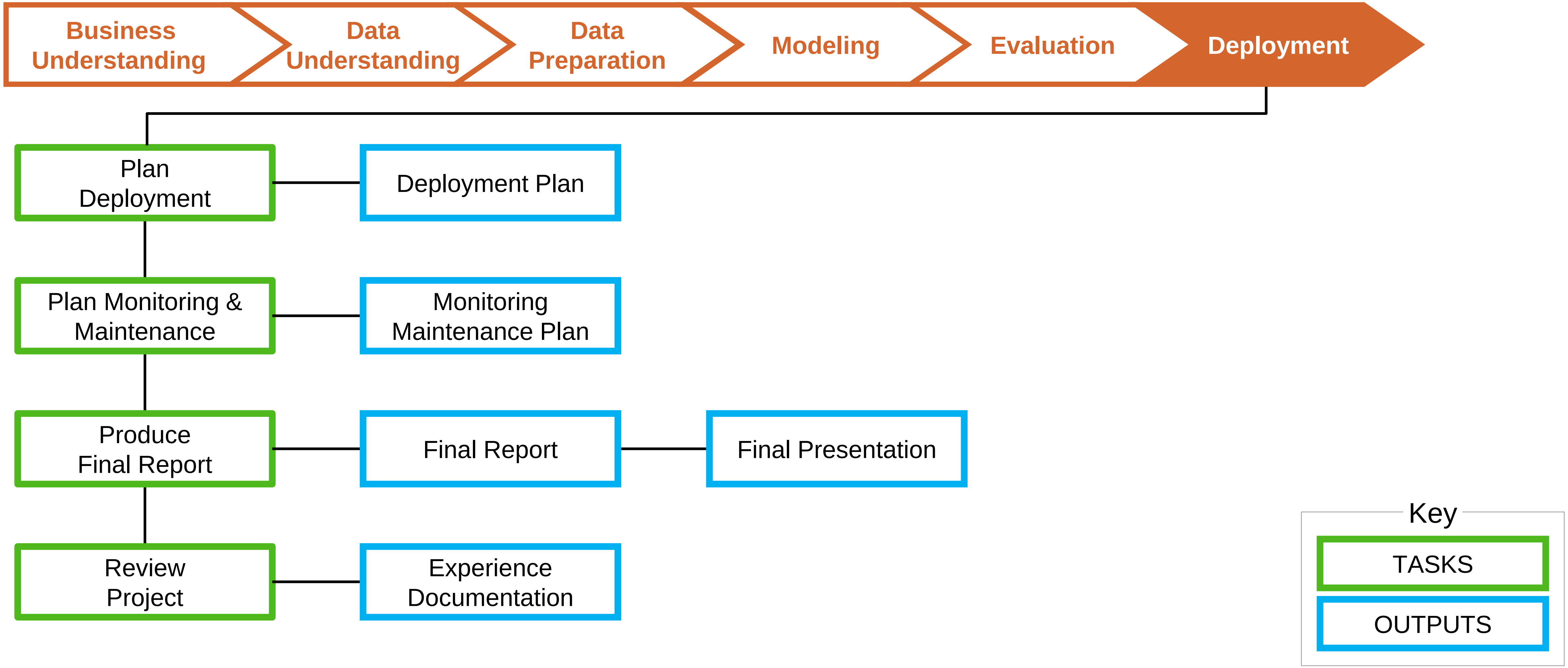
CRISP-DM Project Methodology – Recap

CRISP-DM – Phase 5: Evaluation



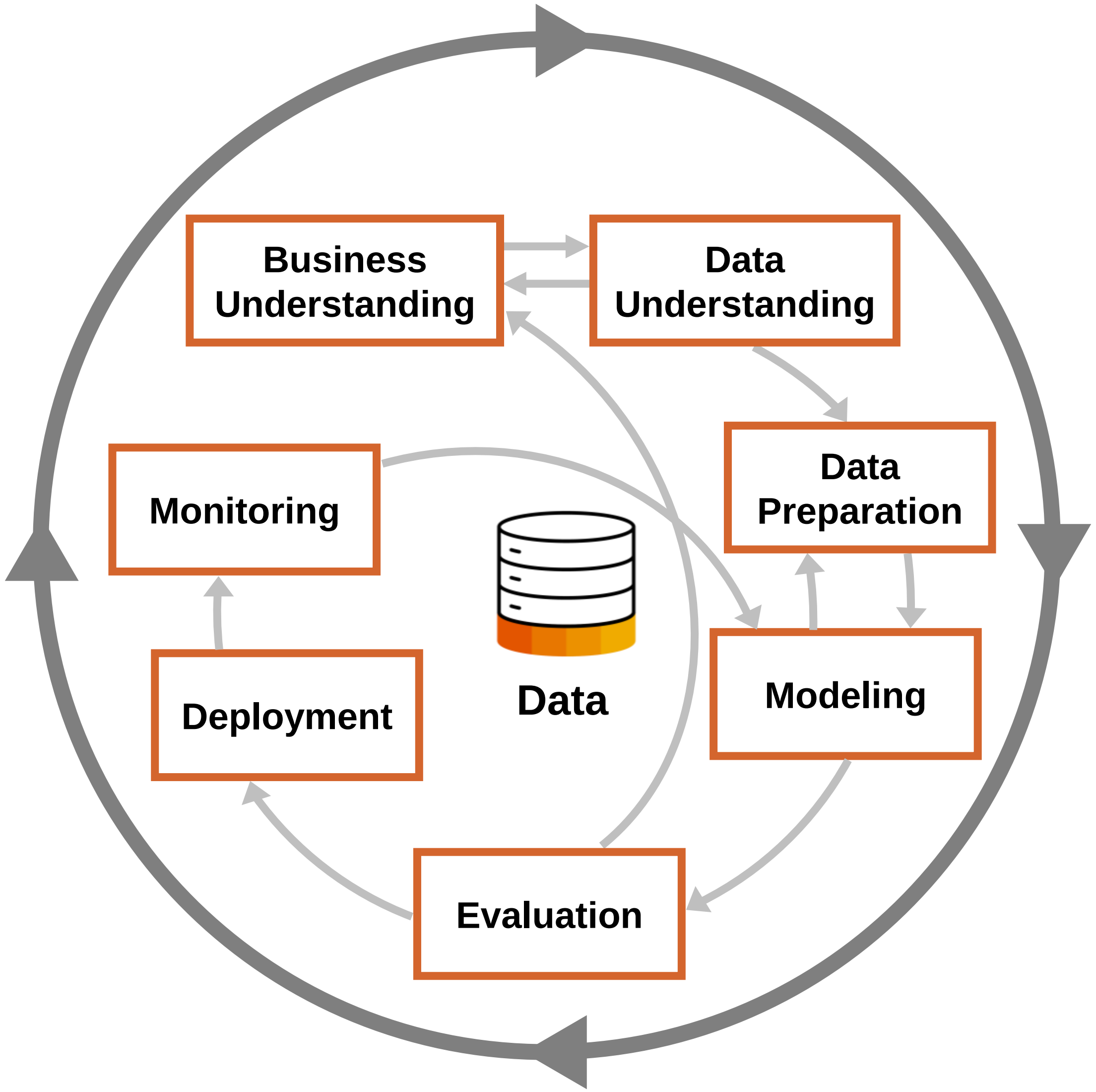
CRISP-DM Project Methodology – Recap

CRISP-DM – Phase 6: Deployment



CRISP-DM Project Methodology – Recap

CRISP-DM – Monitoring phase



CRISP-DM Project Methodology – Recap

Summary

- In this unit, you have examined the 6 generic phases of the CRISP-DM project methodology
- The six phases are business understanding, data understanding, data preparation, modeling, evaluation, and deployment.
- You have also looked briefly at the different tasks that are required in each phase.
- Sometimes, data scientists add in an extra phase to monitor the models, so they are aware when a model's performance degrades and needs updating.
- You will follow this methodology through the next 4 weeks of this course.

Thank you.

Contact information:

open@sap.com

Week 1: Case Study Introduction

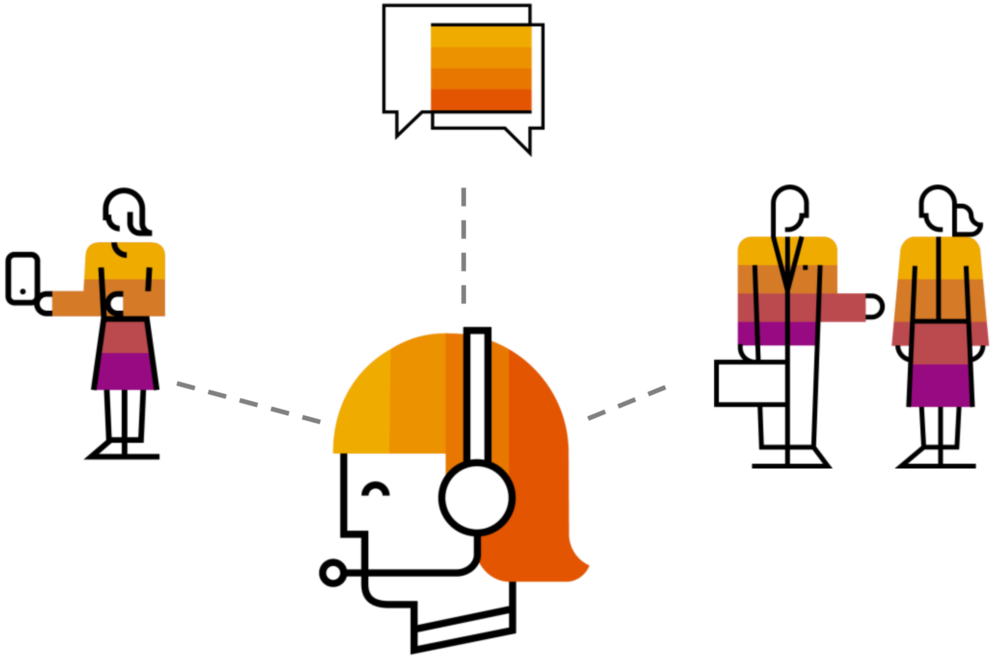
Unit 2: Introduction to the Telco Case Study

Run Simple

openSAP
open.sap.com

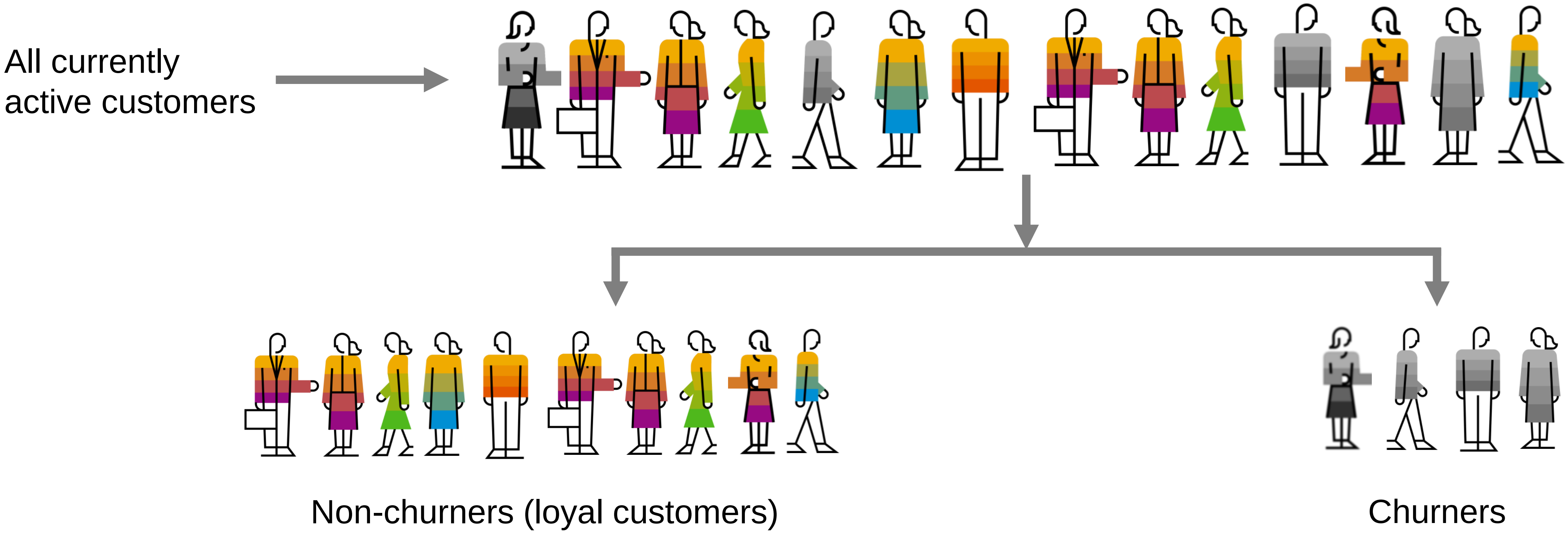
Introduction to the Telco Case Study

Overview



Introduction to the Telco Case Study

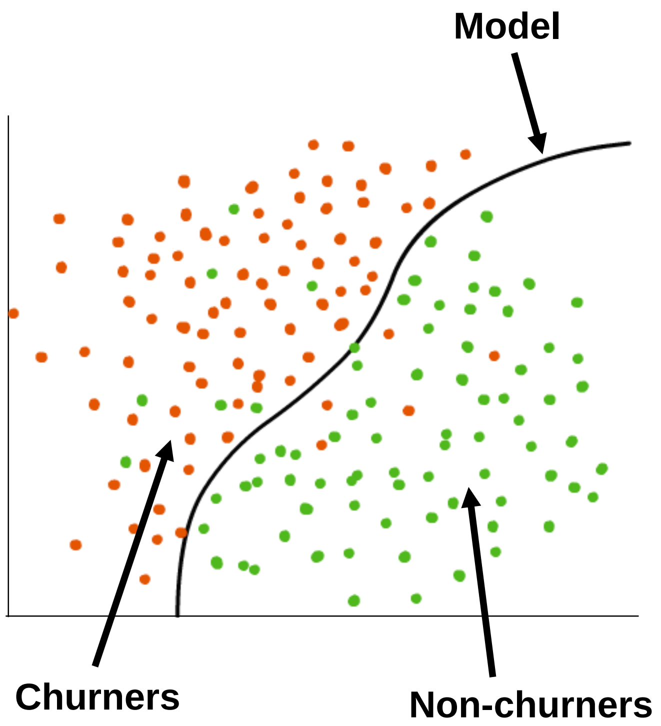
What is a telco churn prediction model?



Introduction to the Telco Case Study

Classification models

- Build a model to describe the attributes of those customers who have not churned, in contrast to those who have churned
- Predict which customers are most likely to churn in future
- Develop strategies to maximize the retention of customers
- The type of model most often used in churn analysis is referred to as a **“classification”** model



https://medium.com/fuzz/machine-learning-classification-models-3040f71e2529

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

Introduction to the Telco Case Study

Classification models in SAP Predictive Analytics automated functionality

- Each input variable is assigned a regression coefficient, b in the equation below.
- The input variables, called “explanatory” variables, are represented by x in the equation below.
- This equation is what we call the “model”.
- The target variable is represented by Y in the equation below.

$$Y = a + b_1 * x_1 + b_2 * x_2 + \dots . b_n * x_n$$

Where

Y is the target **(high values indicate a customer will churn and low values indicate non-churn)**

a is a constant value, defined by the regression algorithm

$b_{1 \text{ to } n}$ are regression coefficients assigned by the regression algorithm

$x_{1 \text{ to } n}$ are the categories of each of the explanatory variables

Regression Analysis:

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

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2845248/

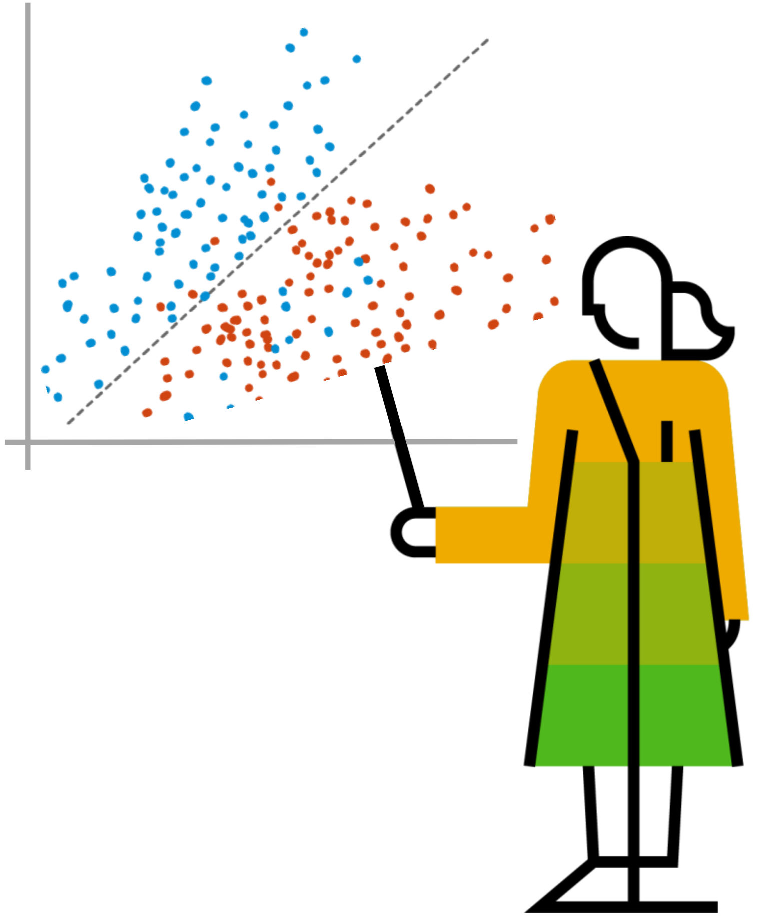
Polynomial:

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

Introduction to the Telco Case Study

Classification models

- We use historical data, where we know the values of the explanatory variables and the value of the target variable.
- The regression process estimates the values of *a* and *b* so that the model estimates the target values as accurately as possible.
- Summing *a* (the constant value) and each (*b* times *x*) multiple gives an estimated value of the target variable Y, which we call a “score”.
- A high score indicates that the customer has a high chance to churn.



Introduction to the Telco Case Study

Explanatory variables

- The “explanatory” variables are usually numeric and categorical and describe the attributes of each customer.
  - In a telco churn model, the explanatory variables represent information about the customer
  - A data scientist will also create a range of “derived” variables

- The overall objective of the model is to differentiate the churners and non-churners.

Typical Telco Explanatory Variables

- Accounts (e.g. tenure, dealer info, SIM info)
- Demographic (e.g. nationality, age)
- Call Centre Info (e.g. number of total complaints)
- Handset (e.g. model)
- Geography (e.g. most called geographical location)
- Usage (e.g. number of inbound/outbound calls)
- Network (e.g. dropped calls)
- Recharge (e.g. amount of first top-up)
- Revenue (e.g. average revenue per user (ARPU))
- Marketing Campaign (e.g. acceptance rate for marketing campaigns)
- etc.

Introduction to the Telco Case Study

Target variable

- In a classification model, the “target” variable in the model build data set is usually coded as a binary variable, i.e. Yes / No or 1 / 0.
- Specifically in a churn model, the target variable is often coded by the data scientist as 1 if the customer churned, or 0 if they did not churn.

Introduction to the Telco Case Study

"Build and apply" in predictive modeling

Model Build

(the Learning Phase)

Predictive models are built or "trained" on historical data with a known outcome.

Model Apply

(the Applying Phase)

Once the model has been built, it is applied onto new, more recent data, which has an unknown outcome (because the outcome is in the future).

Introduction to the Telco Case Study

Model build – Prepare analytical data set (ADS)

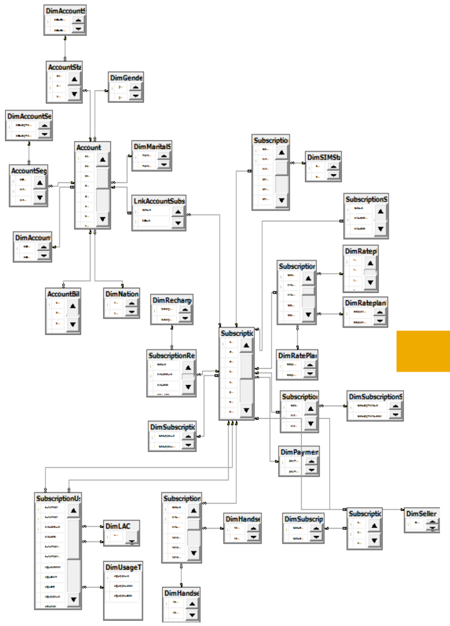
CUSTOMER

USAGE

NETWORK

ACCOUNTS

etc



Date Warehouse

- In this scenario, the user-defined reference date is 2016-03-31.
- All dynamic variables are calculated relative to this date.
- The usage variables are calculated for 3 months prior to the reference date: M0 refers to January 2016, M1 to February, and M2 to March.
- The target represents churn in the period one or two months after the reference date.

		EXPLANATORY VARIABLES													
UNIQUE_ID	USER DEFINED	CUSTOMER			VOICE CALL USAGE (MONTHLY AGGREGATES)						DATA USAGE (MONTHLY AGGREGATES MB)			..	CHURN
LINE_NUMBER	REFERENCE_DATE	AGE_YEARS	GENDER	TENURE_MTHS	CALL_CNT_M0	CALL_CNT_M1	CALL_CNT_M2	CALL_DUR_M0	CALL_DUR_M1	CALL_DUR_M2	DATA_M0	DATA_M1	DATA_M2	..	TARGET
7809702612	2016-03-31	18	Male	4	10	12	24	600	456	669	2406	2406	982	..	1
6139214653	2016-03-31	26	Female	6	27	20	5	556	729	1452	3803	3803	4096	..	0
7809538328	2016-03-31	57	Male	6	5	9	3	789	885	639	2453	2453	4096	..	0
7783183499	2016-03-31	89	Male	9	2	4	12				407	407	281	..	0
7788829560	2016-03-31	34	Female	12	..	..	..	..	..	..	3833	3833	4096	..	1
6132919446	2016-03-31	29	..	9	..	..	..	..	..	..	..	..	..	..	0
4163998288	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
7054925633	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
6047270454	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	1
6134013046	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
7802399721	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

Model Build

Analytical Data Set

Introduction to the Telco Case Study

Model build – Build predictive model

		EXPLANATORY VARIABLES															
UNIQUE_ID	USER DEFINED	CUSTOMER			VOICE CALL USAGE (MONTHLY AGGREGATES)						DATA USAGE (MONTHLY AGGREGATES MB)			..	CHURN
LINE_NUMBER	REFERENCE_DATE	AGE_YEARS	GENDER	TENURE_MTHS	CALL_CNT_M0	CALL_CNT_M1	CALL_CNT_M2	CALL_DUR_M0	CALL_DUR_M1	CALL_DUR_M2	DATA_M0	DATA_M1	DATA_M2	..	TARGET
7809702612	2016-03-31	18	Male	4	10	12	24	600	456	669	2406	2406	982	..	1
6139214653	2016-03-31	26	Female	6	27	20	5	556	729	1452	3803	3803	4096	..	0
7809538328	2016-03-31	57	Male	6	5	9	3	789	885	639	2453	2453	4096	..	0
7783183499	2016-03-31	89	Male	9	2	4	12				407	407	281	..	0
7788829560	2016-03-31	34	Female	12	..	..	..	..	..	..	3833	3833	4096	..	1
6132919446	2016-03-31	29	..	9	..	..	..	..	..	..	..	..	..	..	0
4163998288	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
7054925633	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
6047270454	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	1
6134013046	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
7802399721	2016-03-31	..	..	..	..	..	..	..	..	..	..	..	..	..	0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	0
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

- A high score indicates that the unique ID has a target where CHURN = 1
- A low score indicates that the unique ID has a target where CHURN = 0

Model Build Analytical Data Set

User-Defined Reference Date = 2016-03-31

SAP PREDICTIVE ANALYTICS

Build Classification Model

$$Y = a + b_1 * x_1 + b_2 * x_2 + \dots\ b_n * x_n$$

For example:

Score = 0.5634 + (0.3794 x AGE_YEARS) + (0.159 x TENURE_MTHS) + (0.0456 x CALL_CNT_M0) + ……….)

Introduction to the Telco Case Study

Model apply – Apply model every month

		EXPLANATORY VARIABLES													
UNIQUE_ID	USER DEFINED	CUSTOMER			VOICE CALL USAGE (MONTHLY AGGREGATES)						DATA USAGE (MONTHLY AGGREGATES MB)			..	CHURN
LINE_NUMBER	REFERENCE_DATE	AGE_YEARS	GENDER	TENURE_MTHS	CALL_CNT_M0	CALL_CNT_M1	CALL_CNT_M2	CALL_DUR_M0	CALL_DUR_M1	CALL_DUR_M2	DATA_M0	DATA_M1	DATA_M2	..	TARGET
6132435172	2016-06-30	24	Female	12	3	6	11	120	557	538	337	1146	578	..	0.2856
6132461613	2016-06-30	56	Male	7	12	34	4	248	389	640	2585	2845	2469	..	-0.1945
6132464181	2016-06-30	18	Male	9	26	20	35	319	279	170	228	3700	5618	..	0.0024
6132465666	2016-06-30	22	Female	10	6	7	6				597	256	149	..	1.4896
6132470392	2016-06-30	51	Female	18	..	..	..	..	..	..	3990	259	3890	..	-0.7267
6132470615	2016-06-30	29	..	5	..	..	..	..	..	..	..	..	..	..	0.6678
6132471047	2016-06-30	..	..	..	..	..	..	..	..	..	..	..	..	..	1.0256
6132472127	2016-06-30	..	..	..	..	..	..	..	..	..	..	..	..	..	-0.0049
6132499775	2016-06-30	..	..	..	..	..	..	..	..	..	..	..	..	..	..
6132500423	2016-06-30	..	..	..	..	..	..	..	..	..	..	..	..	..	..
6132510447	2016-06-30	..	..	..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..

Please notice that the score is simply the output from the model equation. It can have negative values, and can have values greater than 1.

It is not a probability.

Model Apply Analytical Data Set

User-Defined Reference Date = 2016-06-30

To apply the model every month, increase reference date + 1 month to update the explanatory variables in the correct time frame.

SAP PREDICTIVE ANALYTICS

Apply Classification Model

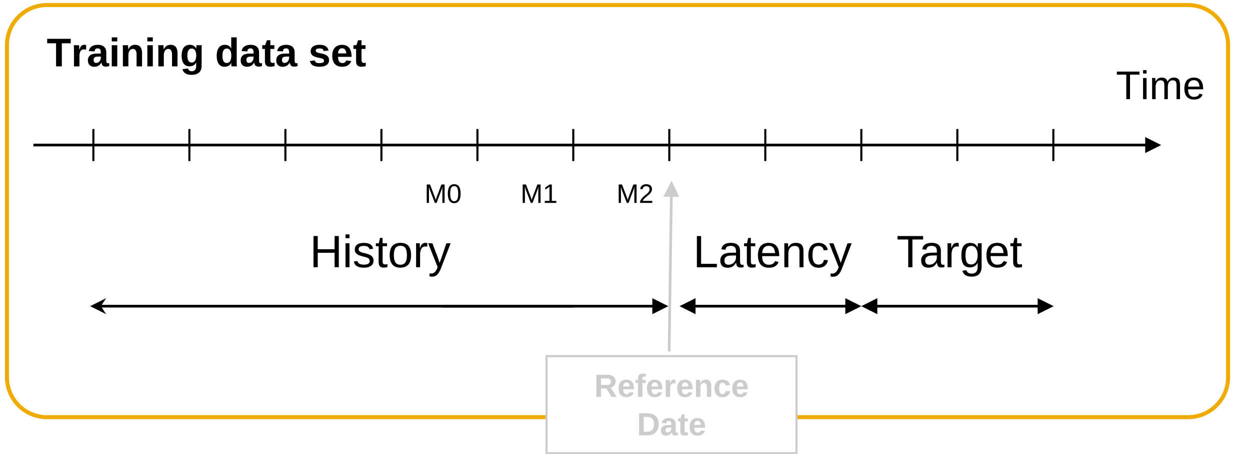
$$Y = a + b_1 * x_1 + b_2 * x_2 + \dots.\ b_n * x_n$$

Apply the model to calculate a score based on all of the explanatory variables for each unique ID.

- A high score indicates that the unique ID has a high potential to churn.
- A low score indicates that the unique ID has a low potential to churn.

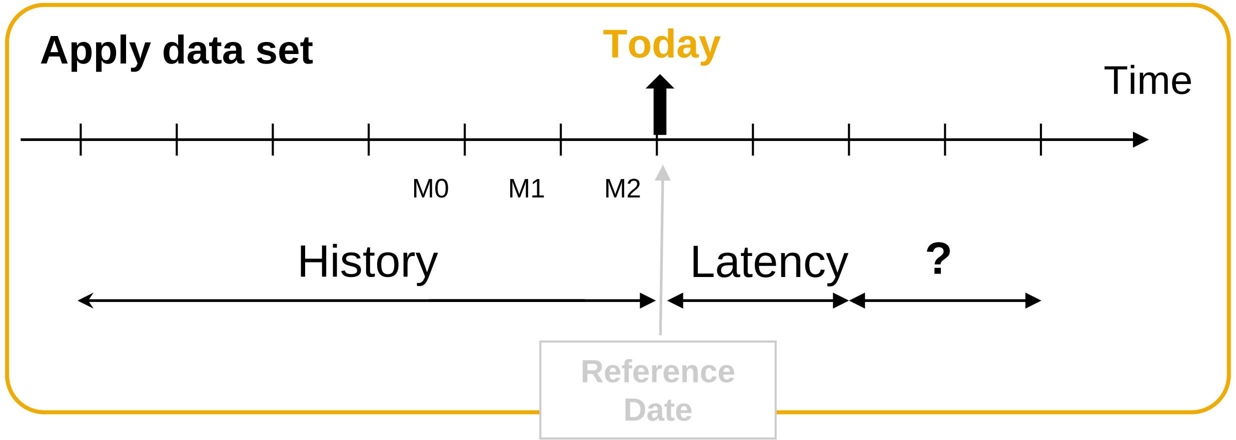
Introduction to the Telco Case Study

Recap – Data-set timeframes and latency periods



For the training data set, the target must be known. It has occurred after the reference date.

For the apply data set, the target is in the future and is therefore unknown.



Introduction to the Telco Case Study

Summary

- You use classification models to predict if a customer will churn or not.
- For predictive churn modeling, data sets can have a history period, a latency period, and a target period. The start and end of these periods are defined by a reference date.
- The model is an equation, and the output from the model is a “score” – a <u>high score</u> indicates that the unique ID has a high potential to churn; a <u>low score</u> indicates that the unique ID has a low potential to churn.
  - The “score” is simply the output from the model equation. It can have negative values, and can have values greater than 1. It is <u>not</u> a probability.
- One of the output options in the automated functionality in SAP Predictive Analytics is to output “probabilities” as well as “scores”. The model scores are mapped into a probability, which varies from 0 to 1. There are no negative probabilities and the maximum value is 1.
- In your churn model, a <u>high probability</u> indicates that the unique ID has a high probability to churn; a <u>low probability</u> indicates that the unique ID has a low probability to churn.

Thank you.

Contact information:

open@sap.com

Week 1: Case Study Introduction

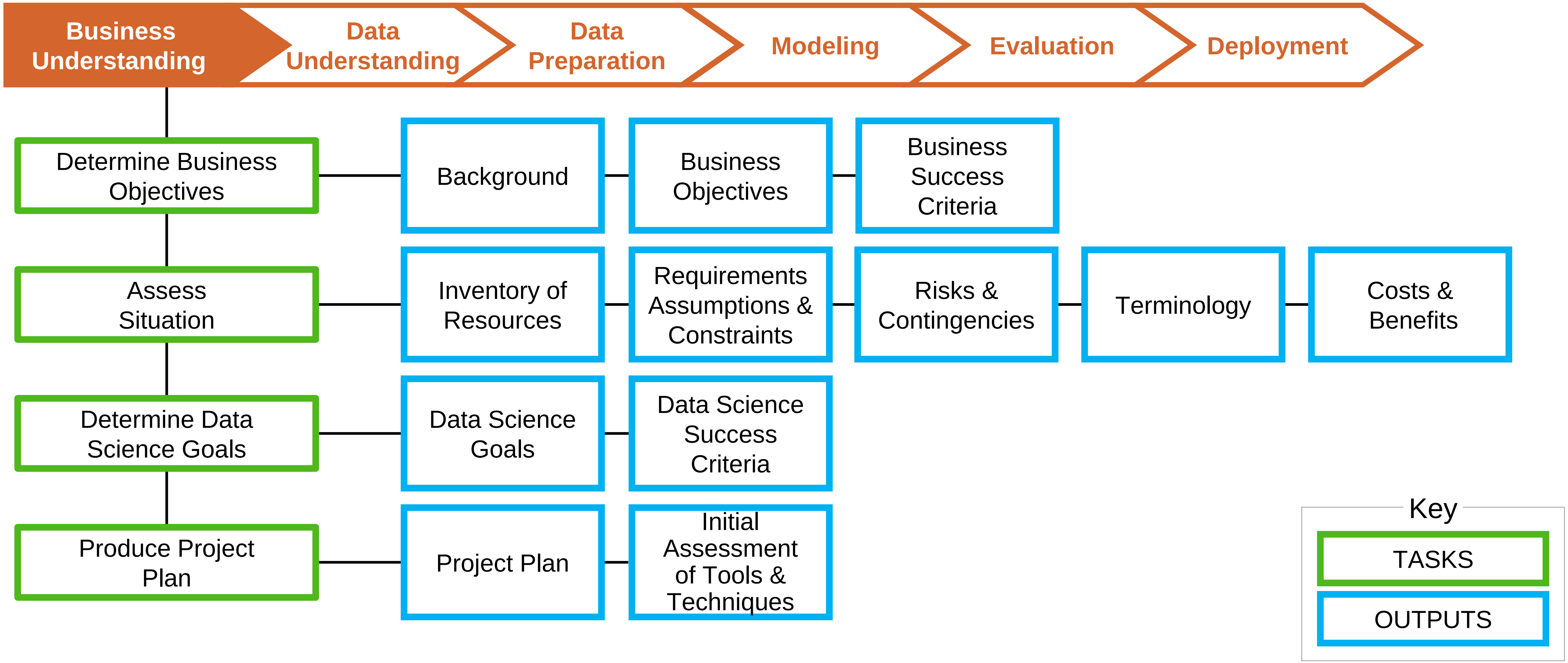
Unit 3: Understanding the Business Requirements

Run Simple

open.sap.com

Understanding the Business Requirements

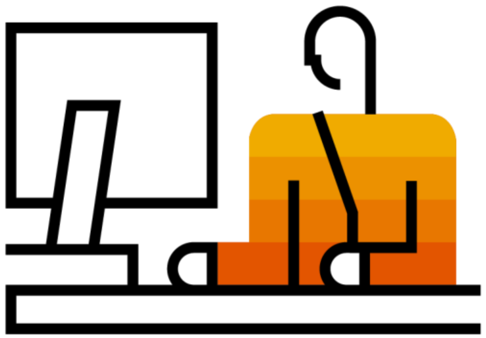
CRISP-DM – Phase 1: Business Understanding



Understanding the Business Requirements

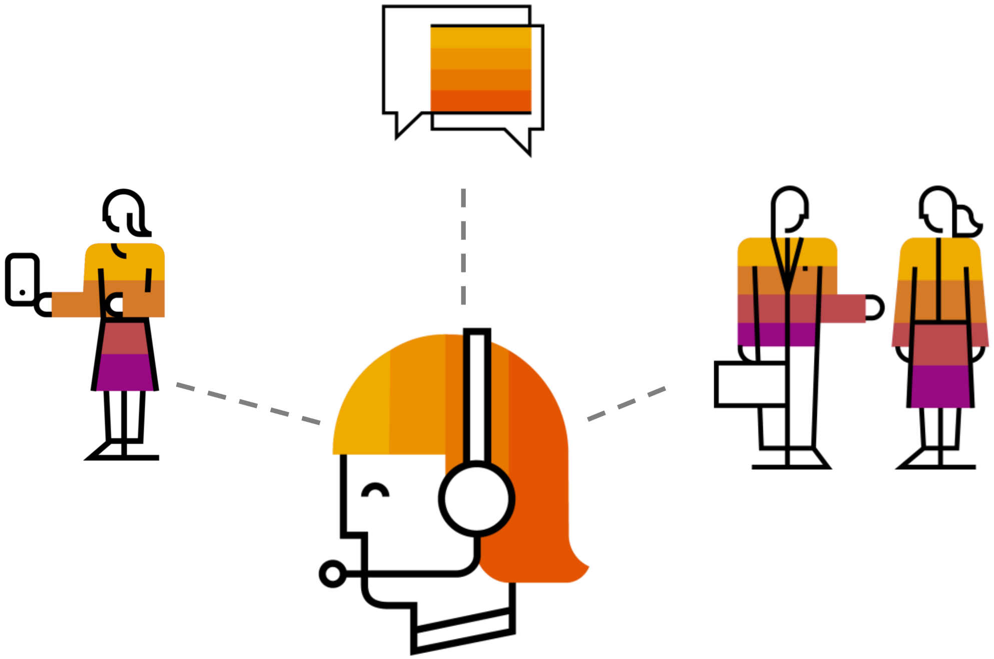
Determine business objectives

- Task
  - The first objective of the data analyst is to thoroughly understand, from a business perspective, what the client really wants to accomplish.
- Outputs
  - Background
  - Business objectives
  - Business success criteria



Understanding the Business Requirements

Determine business objectives – Background



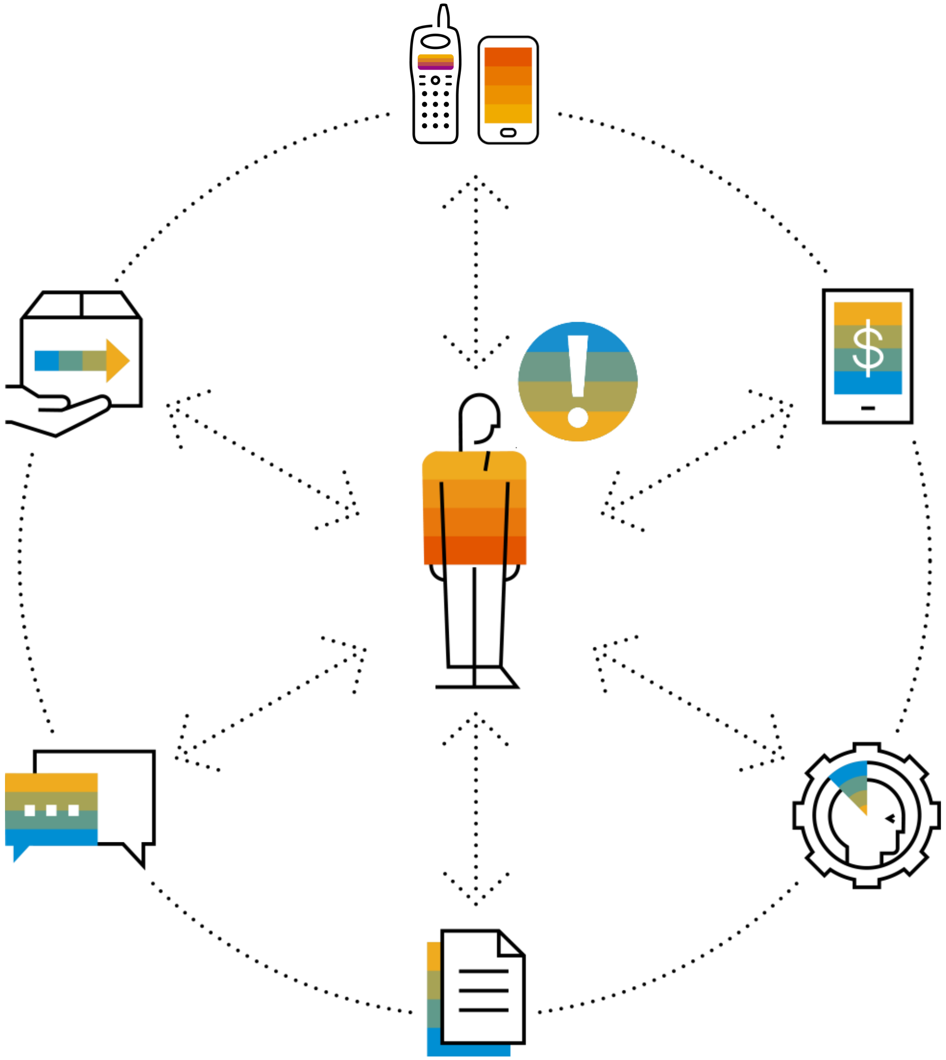
Understanding the Business Requirements

Determine business objectives – Background

- The Premium Service Plan is for “prepaid” customers.
- There is a “bundle” of services provided in this plan:
  - 4GB 4G data
  - 500 local minutes of voice calls
  - Unlimited local texts
- The services in the bundle last for 30 days, then it renews automatically.
- Payment is taken directly from a customer’s bank account.
- If the customer does not have sufficient credit to pay for the service, or if they opt out, then they are classed as **churned**.

Understanding the Business Requirements

Determine business objectives – Background

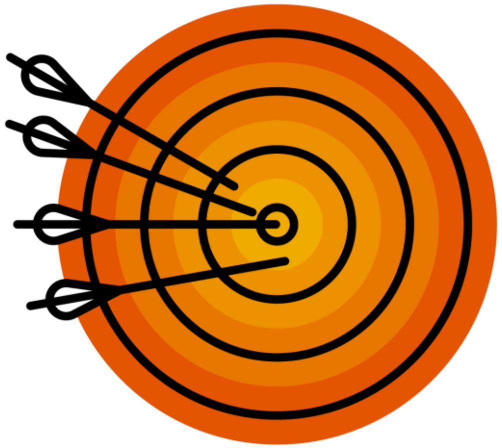


Understanding the Business Requirements

Determine business objectives

Goals:

- Develop a predictive model
- Identify the key contributing factors, or customer characteristics
- Analyze each customer’s social network
- Productionize the model
- Develop a segmentation

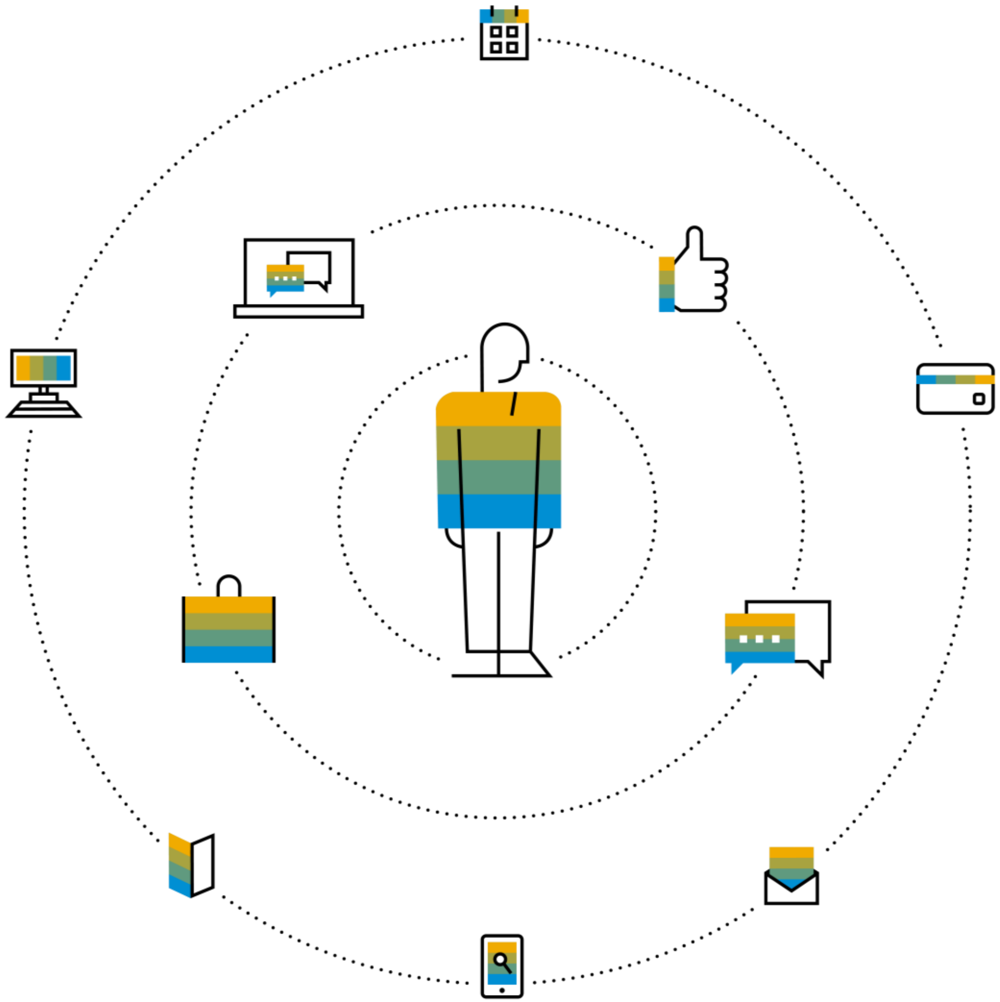


Understanding the Business Requirements

Determine business objectives – Business success criteria

The success factors for the churn model:

- Model accuracy
- Model robustness
- The model must be easy to productionize



Understanding the Business Requirements

Assess situation

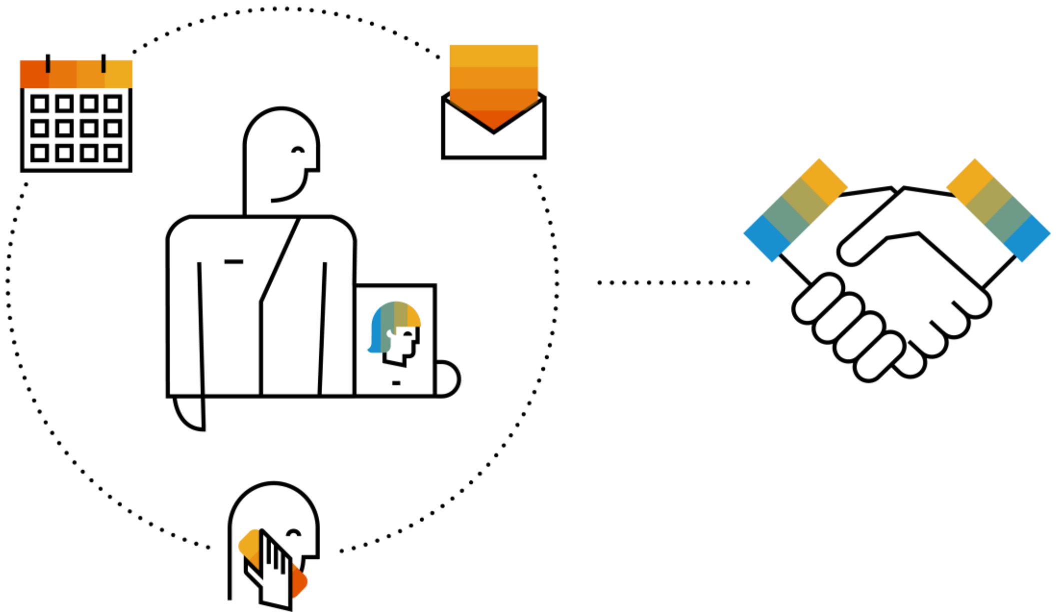
- Task
  - In the previous task, your objective is to quickly get to the crux of the situation. Here, you want to flesh out the details.
- Outputs
  - Inventory of resources
  - Requirements, assumptions, and constraints
  - Risks and contingencies
  - Terminology
  - Costs and benefits

Understanding the Business Requirements

Assess situation – Resources, assumptions, constraints, risks, costs, and benefits

In this task, you flesh out the details of the project.
For example:

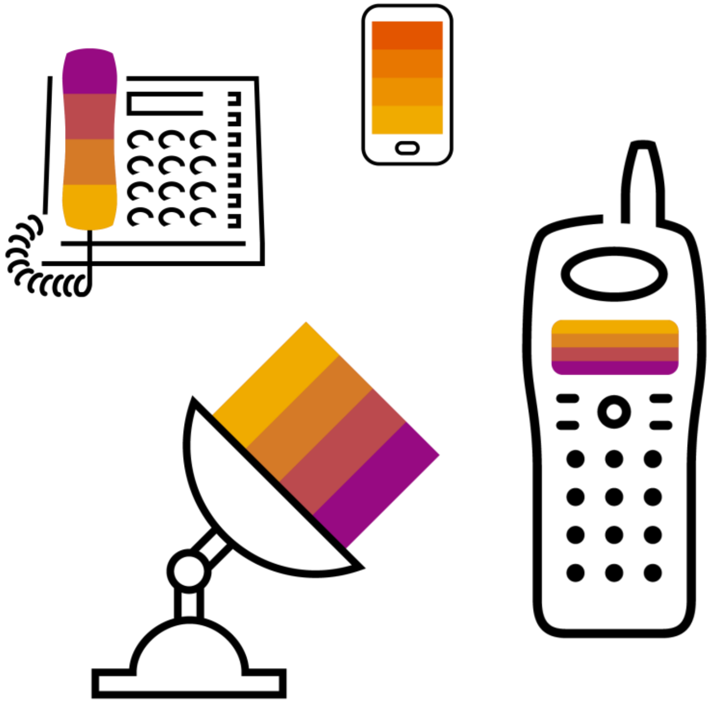
- Inventory of Resources –
  - You are the only analyst assigned to this project.
  - The telco will give you access to a business analyst and a data expert when required.
  - The telco will also supply you with all available data and information about the data.
  - The telco has asked that you use SAP Predictive Analytics automated modeling techniques, because of the quick development time, high accuracy, and ease of use.



Understanding the Business Requirements

Assess situation – Telecommunications industry terminology

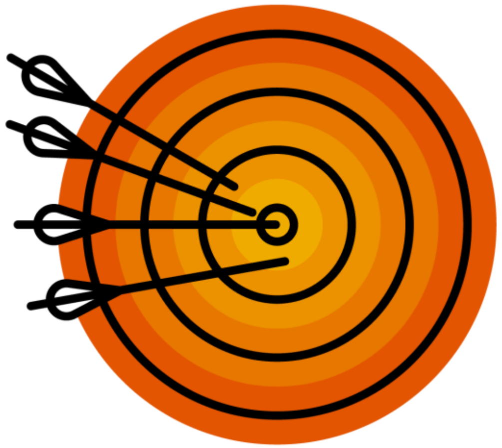
- A “prepaid” mobile phone is a mobile phone for which credit is purchased in advance of service use.
- A “top-up” or “recharge” is where a customer makes a payment to continue to use the service.
- A “bundle” is a mixture of telecommunications services in a single priced product.



Understanding the Business Requirements

Determine data science goals

- Task
  - A *business goal* states objectives in business terminology.
  - A *data science goal* states project objectives in technical terms.
- Outputs
  - Describe data science goals
  - Define data science success criteria



Understanding the Business Requirements

Determine data science goals

The goals of the project from a data science perspective are as follows:

Phase 1

- Develop an initial classification model to predict which customers will churn.
  - History period – 3 months
  - Latency period – 1 month
  - Target period – 1 month

Phase 2

- Develop a social network (link) analysis of the call patterns 1 month prior to the reference date to investigate if this could enhance the churn.

Phase 3

- Develop a k-means cluster model to start to understand more about how customers are using the service.
- This will be a supervised cluster model, and you will use customer spend over the past 3 months as the target variable.

Understanding the Business Requirements

Determine data science success criteria

The success criteria for each model should be agreed with the customer at this early stage.

Phase 1

- Initial churn model
- Success criteria –
  - Predictive power of model to be confirmed with telco when this initial model is developed
  - Prediction confidence of model >= 0.95 so that model is robust
  - Model to be productionized in Predictive Factory

Phase 2

- Develop a social network analysis
- Success criteria –
  - Improved understanding of call patterns that can be used to enhance future predictive models

Phase 3

- Supervised k-means cluster model
- Success criteria –
  - Telco confirms that segment behavior profiles make business sense, and are easy to understand and act on
  - The number of segments should be greater than 3, but less than 10

Understanding the Business Requirements

Produce project plan

- Task
  - Describe the intended plan for achieving the data science goals and thereby achieving the business goals.
- Output
  - Project plan with project stages, duration, resources, etc.
  - Initial assessment of tools and techniques

Understanding the Business Requirements

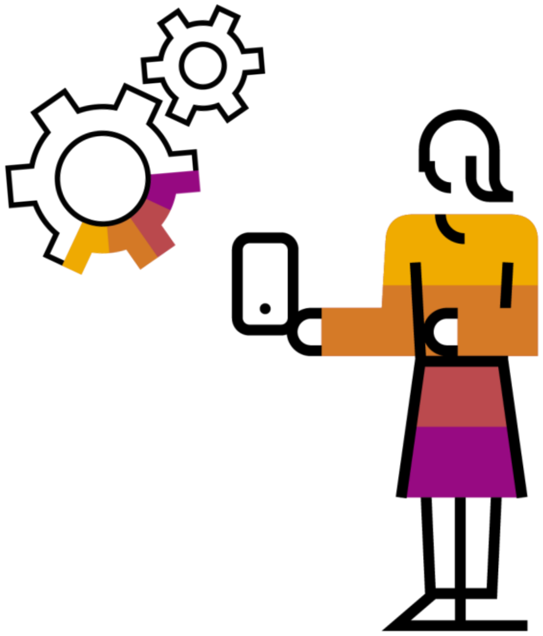
Produce project plan

- This task lists the stages to be executed in the project, together with durations, resources, inputs, outputs, and dependencies.
- For this simple openSAP scenario, there is no need for a project plan. However, here is an example project plan for your information.

		RESOURCES						Project Time Box Example							
		SAP		CUSTOMER				Month 1				Month 2			
	Project Phase	PA	HANA	Business	PA	HANA	0	1	2	3	4	5	6	7	8
Prep Activities	**0 Project Preparation and Readiness**														
	0.1 Infrastructure Readiness														
	0.2 Software Readiness		x			x									
	0.3 Admin Readiness														
Core Activities	**1.0 Business Understanding**														
	1.1 Determine Business Objectives	x		x	x										
	1.2 Assess Situation	x		x	x										
	1.3 Determine Data Science Goals	x		x	x										
	1.4 Produce Project Plan	x		x	x										
	2.0 Data Understanding														
	2.1 Collect Initial Data	x		x	x										
	2.2 Describe Data	x		x	x										
	2.3 Explore Data	x		x	x										
	2.4 Verify Data Quality	x		x	x										
	3.0 Data Preparation														
	3.1 Select Data	x			x										
	3.2 Clean Data	x			x										
	3.3 Construct Data	x			x										
	3.4 Integrate Data	x			x										
	3.5 Format Data	x			x										
	4.0 Modeling														
	4.1 Select Modeling Technique	x			x										
	4.2 Generate Test Design	x			x										
	4.3 Build Model	x			x										
	4.4 Assess Model	x			x										
	5.0 Evaluation														
	5.1 Evaluate Results	x			x										
	5.2 Review Process	x			x										
	5.3 Determine Next Steps	x			x										
	6.0 Deployment														
	6.1 Plan Deployment	x			x										
	6.2 Plan Monitoring & Maintenance	x			x										
	6.3 Produce Final Report	x			x										
	6.4 Review Project	x			x										

Understanding the Business Requirements

Produce project plan – Initial assessment of tools and techniques



- Algorithm
  - A classification algorithm is ideal for the development of a churn model as it will provide the required output – a classification of customers into two groups: churners and non-churners.
  - A classification algorithm will also help identify the important explanatory variables that contribute to the model output.
  - A cluster algorithm, such as k-means, will group customers based on their behavior.

- Tool
  - You will use the SAP Predictive Analytics automated tools because of their ease of use, speed, and accuracy.

Other techniques used by data scientists:
http://www.datasciencecentral.com/profiles/blogs/40-techniques-used-by-data-scientists

k-means:
https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm

Understanding the Business Requirements
Summary

- You have been introduced to the business understanding phase of the project.
- You have determined the business objectives of the project and business success criteria.
- You have also determined the data science goals and data science success criteria.
- You have been asked to develop:
  - a predictive churn model
  - a social network analysis
  - a supervised cluster model
- You have assessed the situation and you will use SAP Predictive Analytics automated technology to build the models.

Thank you.

Contact information:

open@sap.com

Week 1: Case Study Introduction

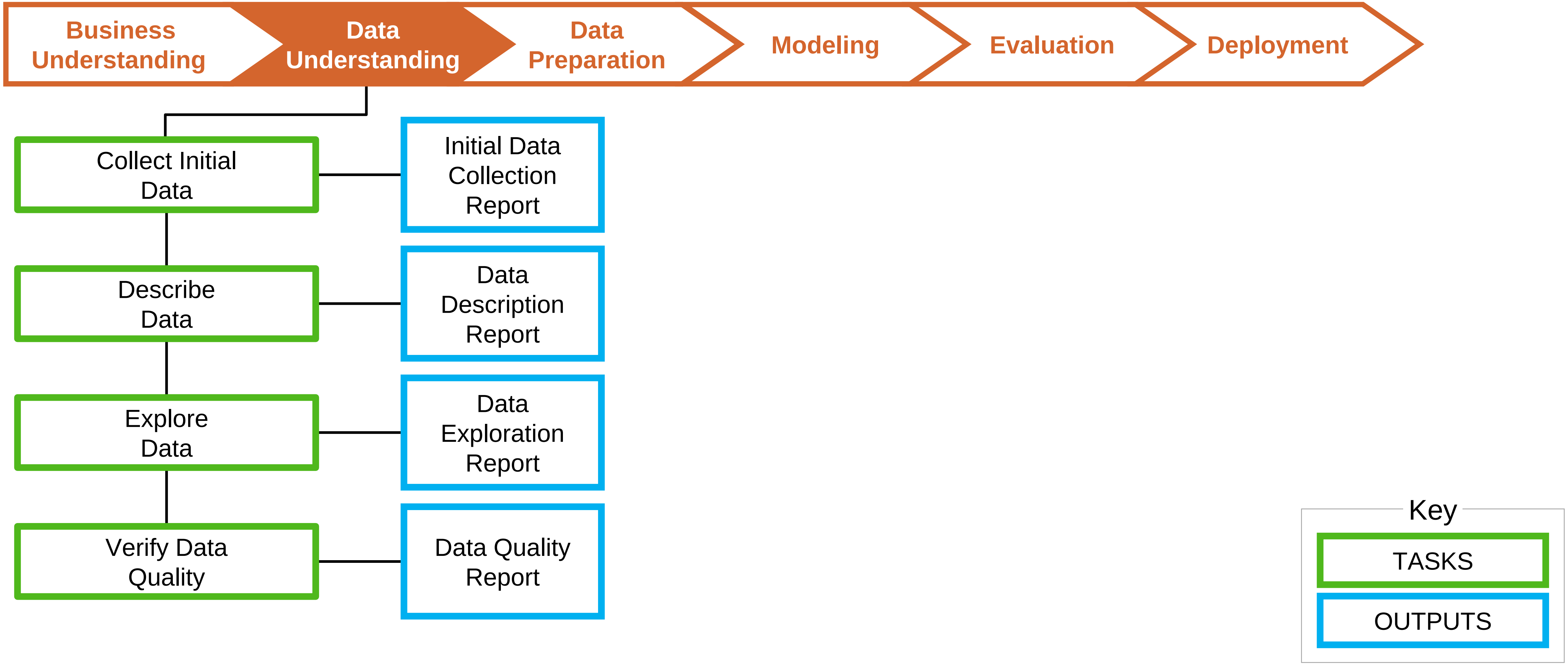
Unit 4: Understanding the Data

Run Simple

openSAP
open.sap.com

Understanding the Data

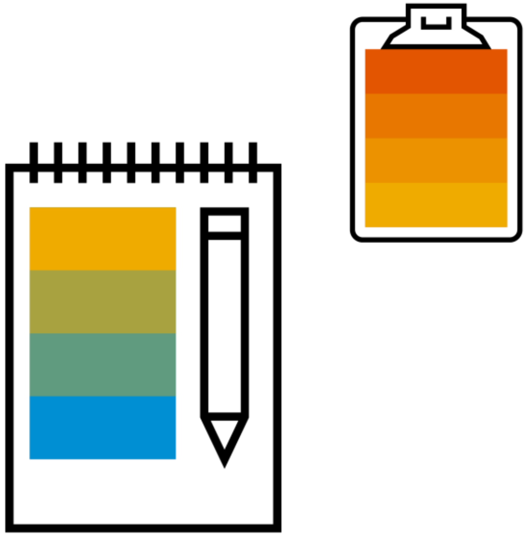
CRISP-DM – Phase 2: Data Understanding



Understanding the Data

Collect initial data

- Task
  - Acquire the data (or access to the data) listed in the project resources.
  - This initial collection includes data loading into the data exploration tool and data integration if multiple data sources are acquired.
- Output – Initial Data Collection Report
  - List the following:
    - The data set (or data sets) acquired
    - The data set locations
    - The methods used to acquire the data sets
    - Any problems encountered
  - Record problems encountered and any solutions



Understanding the Data

Collect initial data

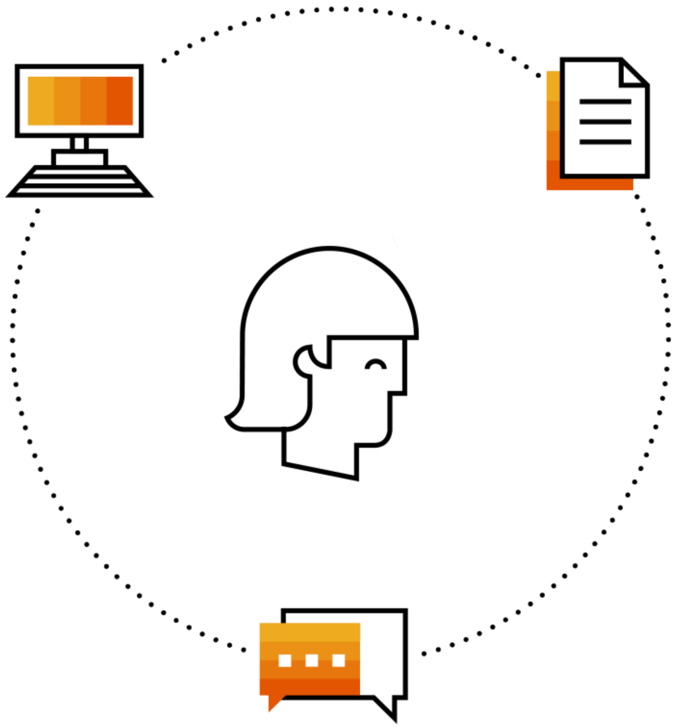
- The telco has made the following data sources available to you:
  - A_NUMBER_FACT
  - CUSTOMER_ID_LOOKUP
  - CUSTOMER
  - CDR
  - DATA_USAGE
  - SPEND_SEGMENTATION

- These data sources are all located in the SAP HANA DB.
- You will be given access information.

Understanding the Data

Describe data

- Task
  - Examine the “surface” properties of the acquired data and report on the results.
- Output – Data Description Report
  - Describe the data that has been acquired, including:
    - The format of the data
    - The quantity of data, e.g. the number of records and fields in each table
    - The identities of the fields
    - Any other surface features of the data that have been discovered



Understanding the Data

Describe data

A_NUMBER_FACT

- This is a list of the unique line numbers (A_NUMBER) associated with each account.
- There are no duplications.
- It is the fact table for the data manipulation, and the other tables used in the analysis can be merged and aggregated to it.
- This is the “entity” in our analysis – it is the object of interest. The goal of the analysis is to identify which A_NUMBERs are going to churn.
- There is 1 column of data.
- There are 7445 rows of data.
- All of the customers have been customers for a minimum of 6 months, and were classed as active (i.e. they had not churned) as of the end of March.

	A_NUMBER
1	2042930441
2	2502048322
3	2502164353
4	2502280241
5	2503072523
6	2503383993
7	2504153759
8	2504159954
9	2504866064
10	2505070225
11	2505162314
12	2505165087
13	2505897474
14	2506182840
15	2506187882
16	2506192841
17	2506197212
18	2507278441
19	2507440086
20	2507514885
21	2507972669

Understanding the Data

Entity

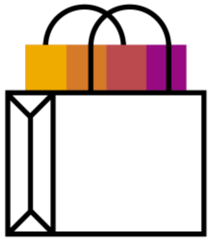
With any predictive modeling, you will need to identify the “entity” for the analysis.

- An entity is the object targeted by the model.
- It may be a customer, a product, or a store, etc., and is usually identified by a unique identifier.
- The entity defines the granularity of the analysis.

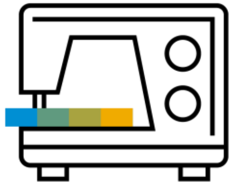
Items of significance to an enterprise are data entities

Sale

Customer



Material



Product

Understanding the Data

Describe data

CUSTOMER_ID_LOOKUP

- This is a lookup table that links a CUSTOMER_ID to the A_NUMBER.
- There are 2 columns of data.
- Statistical analysis shows there are 7445 rows of data.

	A_NUMBER	CUSTOMER_ID
1	2042930441	1000172
2	2502048322	1000198
3	2502164353	1000210
4	2502280241	1000213
5	2503072523	1000258
6	2503383993	1000260
7	2504153759	1000261
8	2504159954	1000277
9	2504866064	1000303
10	2505070225	1000313
11	2505162314	1000329
12	2505165087	1000356
13	2505897474	1000365
14	2506182840	1000366
15	2506187882	1000382
16	2506192841	1000384
17	2506197212	1000394
18	2507278441	1000403
19	2507440086	1000408
20	2507514885	1000410
21	2507972669	1000418

Understanding the Data

Describe data

CUSTOMER

	CUSTOMER_ID	GENDER	AGE	ZIP_CODE	DISTRIBUTION_CHANNEL_ID	DEVICE_BRAND_NAME	DEVICE_MODEL_NAME	TENURE_MTHS
1	1000111	M	40	91706	SMO0001	OnePlus	3T	7
2	1000112	M	62	49509	AUC0001	Google	Pixel	10
3	1000113	F	53	11213	PDS0001	Apple	iPhone 7	10
4	1000114	M	15	91335	WMN00001	Google	Pixel XL	15
5	1000115	F	48	70560	SPR00001	Google	Pixel	6
6	1000116	F	55	90650	WLM0001	Apple	iPhone 7	14
7	1000117	M	21	90805	PDS0001	Apple	iPhone 7	6
8	1000118	F	29	60623	SMO0001	Apple	iPhone 7	12
9	1000119	F	57	23602	PH00001	Apple	iPhone 7	12
10	1000120	M	32	92647	PDS0001	Apple	iPhone 7	13
11	1000121	M	22	44107	WHP00001	Apple	iPhone 7	11
12	1000122	M	41	92805	XCLL001	Apple	iPhone 7	16
13	1000123	M	51	94565	SPR00001	OnePlus	3T	11
14	1000124	M	38	49017	CRPH0001	Samsung	Galaxy S7 Edge	14
15	1000125	F	21	48205	MH00001	Google	Pixel	10
16	1000126	M	15	90660	PDS0001	Apple	iPhone 7	11
17	1000127	F	41	10466	WLM0001	Huawei	Honor 8	14

- This table contains customer data.
- For each CUSTOMER_ID there is information about the customer’s gender, age, location (ZIP_CODE), distribution channel, handset (DEVICE_BRAND_NAME and DEVICE_MODEL_NAME), and the number of months they have been a customer (TENURE_MTHS).
- There are 8 columns of data.
- Statistical analysis shows there are 7445 rows of data.

Understanding the Data

Describe data

CDR

- This table contains the call detail record (CDR).
- KxIndex is a row number, and can be ignored.
- The data shows the A_NUMBER contacting the B_NUMBER, the TYPE of call (either MMS, VOICE or SMS), the DURATION of the call (only for VOICE calls, in seconds), and the date and time of the call.
- There are 5 columns of data.
- There are 466080 rows of data.

	KxIndex	A_NUMBER	B_NUMBER	TYPE	DURATION	DATE
1	240039	6048666626	6136780178	MMS	0	2016-03-21 09:40:00
2	240040	6478895621	6136398417	SMS	0	2016-03-21 09:40:00
3	240041	6043741234	6049618135	MMS	0	2016-03-21 09:38:00
4	240042	6047900465	6047678042	MMS	0	2016-03-21 09:45:00
5	240043	6472868652	7802787879	VOICE	320	2016-03-21 09:10:00
6	240044	6132332229	6043070495	SMS	0	2016-03-21 09:50:00
7	240045	4165200344	6133634183	VOICE	128	2016-03-21 09:54:00
8	240046	6045915026	6049289905	SMS	0	2016-03-21 09:42:00
9	240047	4167265920	4167236866	VOICE	32	2016-03-21 09:51:00
10	240048	7059437818	4168784898	VOICE	263	2016-03-21 09:07:00
11	240049	7789997999	6472929835	MMS	0	2016-03-21 10:32:00
12	240050	7057167515	4036898500	VOICE	83	2016-03-21 10:51:00
13	240051	6472859001	6135490866	VOICE	228	2016-03-21 10:18:00
14	240052	6132610277	6047156666	MMS	0	2016-03-21 10:20:00
15	240053	6049024843	4167166391	SMS	0	2016-03-21 10:45:00
16	240054	6043384443	6048792558	VOICE	349	2016-03-21 10:28:00
17	240055	6043391418	6137622077	VOICE	112	2016-03-21 10:29:00
18	240056	6043759167	6049610552	SMS	0	2016-03-21 10:53:00
19	240057	7053215622	6044186440	VOICE	5	2016-03-21 10:09:00
20	240058	6049061406	7809082101	SMS	0	2016-03-21 10:13:00
21	240059	6134479383	4168584360	SMS	0	2016-03-21 10:13:00

Understanding the Data

Describe data

DATA_USAGE

- This table contains the data usage for each A_NUMBER, from January through to June 2016, and the percentage of the usage relative to the total data allowance per month.
- The data also contains a flag that indicates if the line number churned in May or June. This is the **target** for the models you will build.
- There are 20 columns of data.
- There are 7445 rows of data.

	A_NUMBER	SERVICE_...	SERVICE_...	Data_Up_...	Voice_Allo...	SMSAllow...	JAN_Data...	JAN_Data...	FEB_Data...	FEB_Data...	MAR_Data...	MAR_Data...	APR_Data...	APR_Data...	MAY_Data...	MAY_Data...	JUN_Data...	JUN_Data...	CHURN_M...	CHURN_J...
1	2042930441	Premium S...	Data Premi...	4096	250	Unlimited	2406	58.759999...	2126	51.920000...	982	24	2420	59.100000...	0	0	0	0	1	0
2	2502048322	Premium S...	Data Premi...	4096	250	Unlimited	3803	92.859999...	4096	100	4096	100	2970	72.530000...	0	0	0	0	1	0
3	2502164353	Premium S...	Data Premi...	4096	250	Unlimited	2453	59.909999...	4096	100	4096	100	2890	70.569999...	0	0	0	0	1	0
4	2502280241	Premium S...	Data Premi...	4096	250	Unlimited	407	9.9600000...	268	6.54	281	6.87	444	10.85	268	6.54	0	0	0	1
5	2503072523	Premium S...	Data Premi...	4096	250	Unlimited	3833	93.599999...	4096	100	4096	100	3072	75	0	0	0	0	1	0
6	2503383993	Premium S...	Data Premi...	4096	250	Unlimited	3576	87.310000...	4096	100	3768	92	3235	79	0	0	0	0	1	0
7	2504153759	Premium S...	Data Premi...	4096	250	Unlimited	3549	86.659999...	2124	51.869999...	2422	59.140000...	3551	86.689999...	2124	51.869999...	0	0	0	1
8	2504159954	Premium S...	Data Premi...	4096	250	Unlimited	3378	82.489999...	2263	55.259999...	2602	63.549999...	3645	88.989999...	2263	55.259999...	2602	63.549999...	0	0
9	2504866064	Premium S...	Data Premi...	4096	250	Unlimited	1891	46.189999...	2499	61.009999...	2349	57.350000...	1823	44.520000...	2499	61.009999...	2349	57.350000...	0	0
10	2505070225	Premium S...	Data Premi...	4096	250	Unlimited	1606	39.219999...	4096	100	4096	100	1885	46.030000...	0	0	0	0	1	0
11	2505162314	Premium S...	Data Premi...	4096	250	Unlimited	1526	37.259999...	4096	100	4096	100	1355	33.090000...	0	0	0	0	1	0

Understanding the Data

Describe data

SPEND_SEGMENTATION

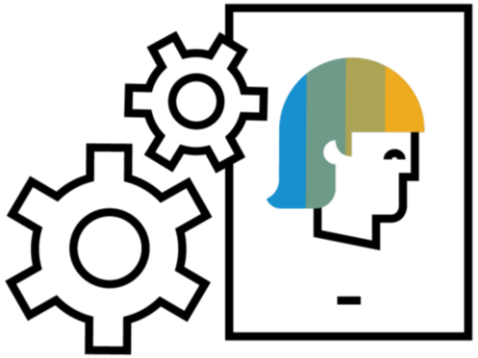
- This table contains the spend for each A_NUMBER over the past 3 months.
- This spend data will be merged to the data set used to build the churn model, and then it can be used as a target in the spend cluster model.
- There are 7445 rows of data.

1	A_NUMBER	SPEND_3_MTHS
2	4036088626	110.92
3	6045624032	197.6
4	6048268415	179.52
5	6047224806	233.5
6	4165681180	224.18
7	6132232939	219.57
8	6043071083	107.32
9	6136398989	114.93
10	6046718265	235.78
11	7789986750	240.3
12	6472868206	207.85
13	7093512622	196.58
14	6048821996	231.57
15	6474078856	221.93
16	6046877221	103.85
17	4033932668	219.53
18	6133559810	183.77
19	6134493715	184.92
20	6135263043	213.9
21	6472084787	189.47

Understanding the Data

Explore data

- Task
  - This task tackles the data questions, which can be addressed using querying, visualization, and reporting.
- Output – Data Exploration Report
  - Describe results of this task, including:
    - First findings or initial hypothesis and their impact on the remainder of the project
    - If appropriate, include graphs and plots



Understanding the Data

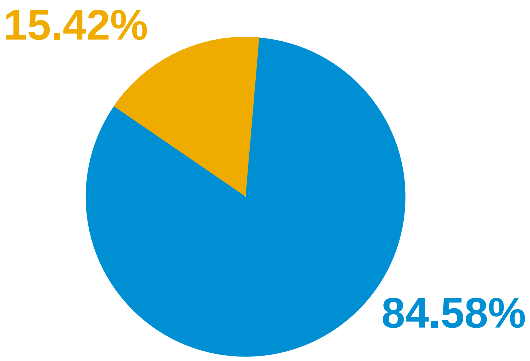
Explore data

The distribution of the continuous variables are shown in the statistical reports. For example:

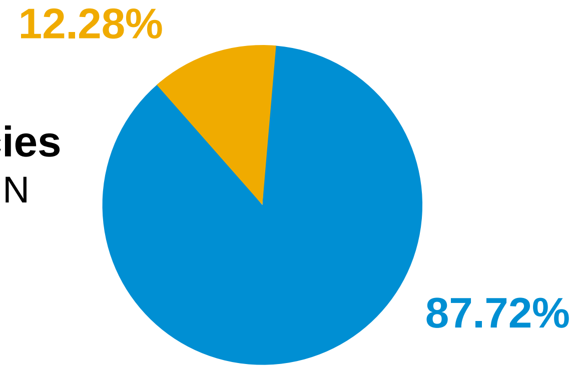
Variable	Min	Max	Mean	Standard Deviation
A_NUMBER	2042930441	9059286491	6,294,251,383.521	1,174,355,100.613
JAN_Data_Usage_MB	99	4096	2,852.387	779.658
JAN_Data_Usage_PCT	2.4399999999999999	100	69.65	19.034
FEB_Data_Usage_MB	4	4096	2,762.401	1,070.006
FEB_Data_Usage_PCT	0.11	100	67.451	26.12
MAR_Data_Usage_MB	24	4096	2,642.25	886.338
MAR_Data_Usage_PCT	0.59999999999999998	100	64.52	21.638
APR_Data_Usage_MB	81	4096	2,740.349	780.839
APR_Data_Usage_PCT	1.99	100	66.915	19.063
MAY_Data_Usage_MB	0	4096	2,331.535	1,371.69
MAY_Data_Usage_PCT	0	100	56.93	33.489
JUN_Data_Usage_MB	0	4096	1,912.334	1,366.594
JUN_Data_Usage_PCT	0	100	46.696	33.368

One important analysis is to ensure there are only two categories in the target variables, with no missing values:

Category Frequencies
Variable: CHURN_MAY



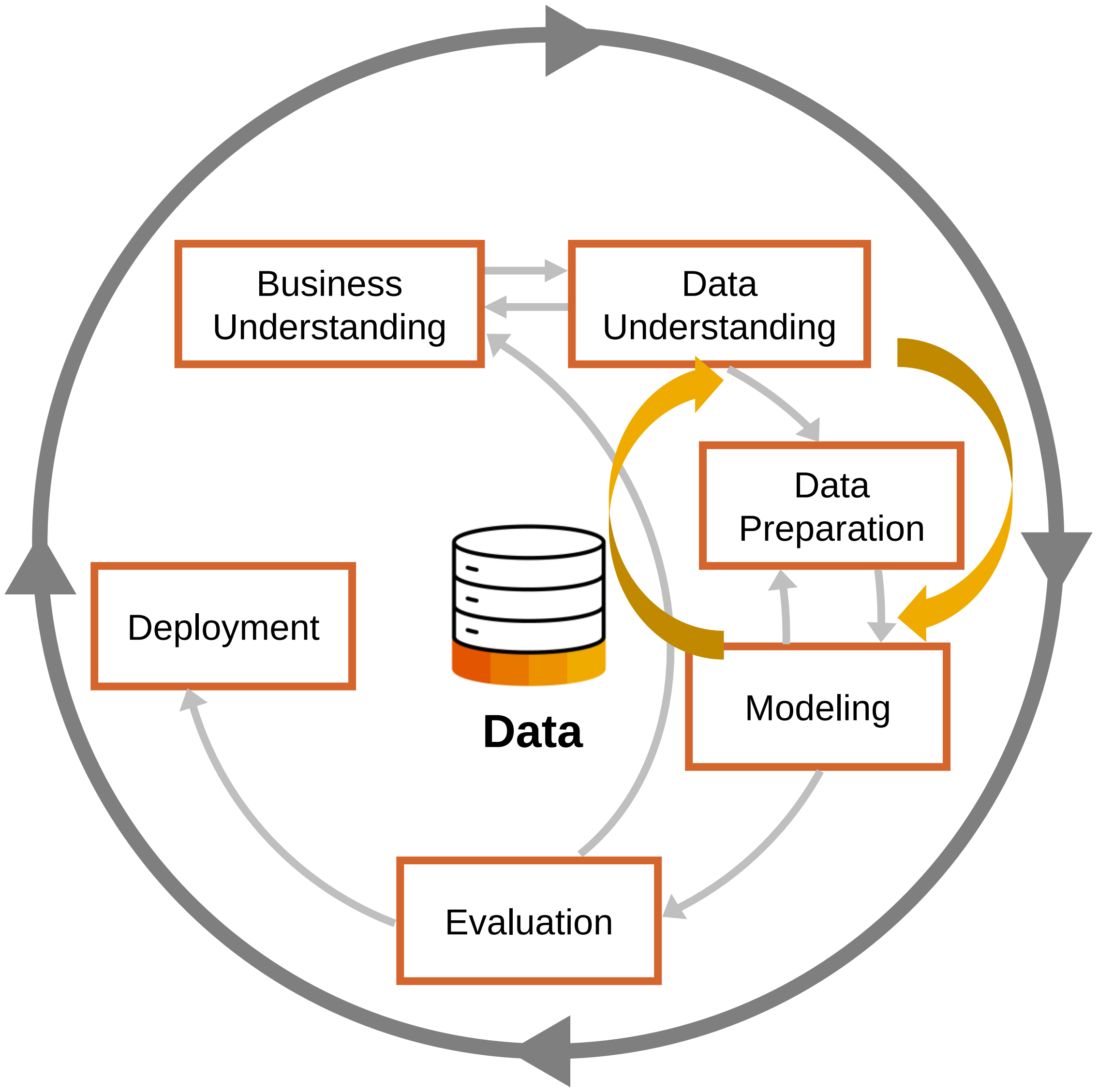
Category Frequencies
Variable: CHURN_JUN



Understanding the Data

Building models first

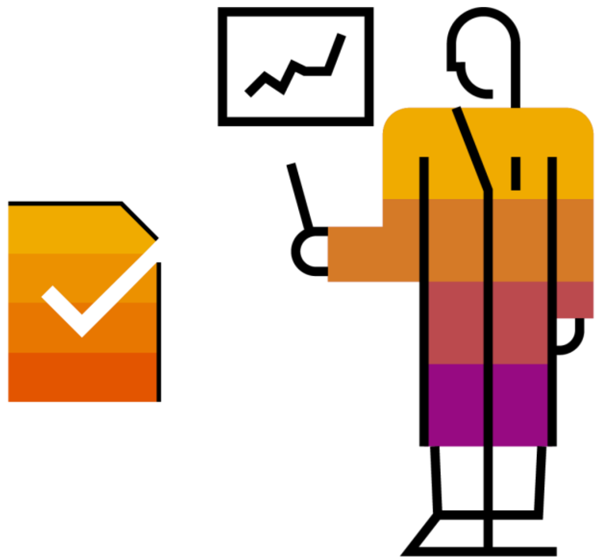
- CRISP_DM is a useful guide, but sometimes there are advantages if you deviate a little.
- You can consider building initial models before the data preparation and data understanding phases have been completed.
- The model will automatically produce a wide range of descriptive statistics, such as cross tabulations of each explanatory variable with the target, and correlations between the explanatory variables.



Understanding the Data

Verify data quality

- Task
  - Examine the quality of the data, addressing questions such as:
    - Is the data complete?
    - Is it correct, or does it contain errors?
    - Are there missing values in the data?
- Output – Data Quality Report
  - List the results of the data quality verification
  - If quality problems exist, list possible solutions



Understanding the Data

Verify data quality – Missing values

- The statistical analysis in Predictive Analytics provides a list of the variables, the value and storage, a count of any missing values, and a row count.
- For example, for the DATA_USAGE table the statistical analysis provides the following information:

Variable	Value	Storage	Missing Count
A_NUMBER	continuous	integer	0
SERVICE_TYPE	nominal	string	0
SERVICE_NAME	nominal	string	0
Data_Up_Allowance_MB	nominal	integer	0
Voice_Allowance_Minutes	nominal	integer	0
SMSAllowance_Num_Messages	nominal	string	0
JAN_Data_Usage_MB	continuous	integer	0
JAN_Data_Usage_PCT	continuous	number	0
FEB_Data_Usage_MB	continuous	integer	0
FEB_Data_Usage_PCT	continuous	number	0
MAR_Data_Usage_MB	continuous	integer	0
MAR_Data_Usage_PCT	continuous	number	0
APR_Data_Usage_MB	continuous	integer	0
APR_Data_Usage_PCT	continuous	number	0
MAY_Data_Usage_MB	continuous	integer	0
MAY_Data_Usage_PCT	continuous	number	0
JUN_Data_Usage_MB	continuous	integer	0
JUN_Data_Usage_PCT	continuous	number	0
CHURN_MAY	nominal	integer	0
CHURN_JUN	nominal	integer	0

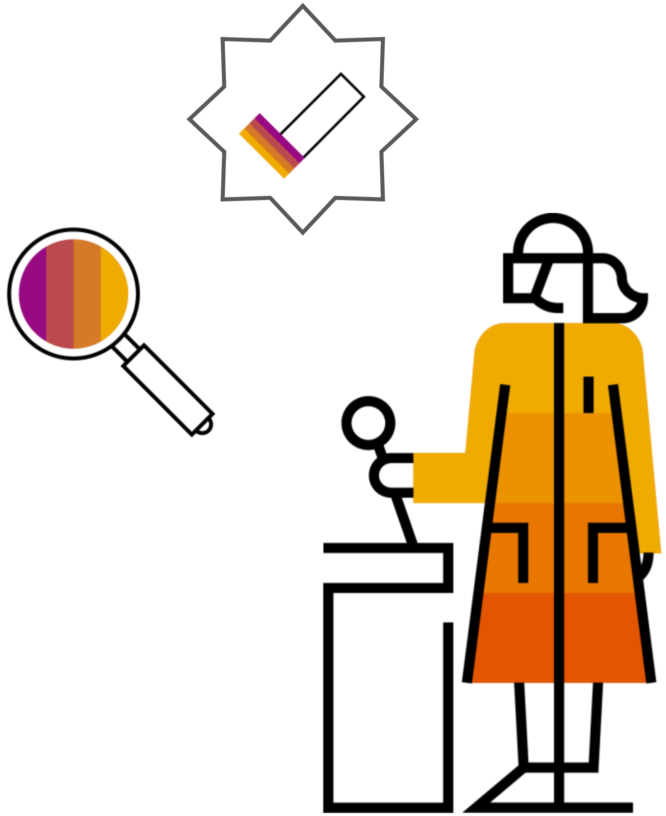
Row Count: **7,445**

Edit Settings

Understanding the Data

Verify data quality

- The SAP Predictive Analytics automated modeling tool provides automated data encoding strategies that deal with missing values and outliers.
- Missing values, outliers, and obvious inconsistencies can be identified in frequency charts and the continuous variable distribution reports.
- They will also be very obvious when you have run an initial test model and examine the model data statistics.



Understanding the Data
Summary

- You have looked at the Data Understanding phase of the project.
- You have accessed and examined the data that is available in the SAP HANA database.
- You have described the data, started to explore and verify the data, and started to check data quality.
- You have used SAP automated analytics to create summary statistics for the tables, and you have seen how to use the output to check the data frequency charts, check for missing values, and produce the statistics for continuous variables.

Understanding the Data

Interesting reading

Data distributions, see https://en.wikipedia.org/wiki/Normal_distribution and https://www.mathsisfun.com/data/standard-normal-distribution.html

Standard deviation, see https://www.mathsisfun.com/data/standard-normal-distribution.html

Correlation, see https://www.mathsisfun.com/data/correlation.html

Leaker variables, see https://www.kaggle.com/wiki/Leakage.

Thank you.

Contact information:

open@sap.com