An Introduction to the CRISP DM methodology

John McConnell – Services and Implementation

Rachel Clinton – Business Development

FAQ's

- Is this session being recorded? Yes, you will get a link to re watch and share the session
- Can I get a copy of the slides? You can access a link so you can download them after the event
- Can we arrange a re-run for colleagues? Yes, just ask us.
- How can I ask questions? All lines are muted so please use the chat facility – if we run out of time we will follow up with you.

Predictive Analytics
for Smarter Business

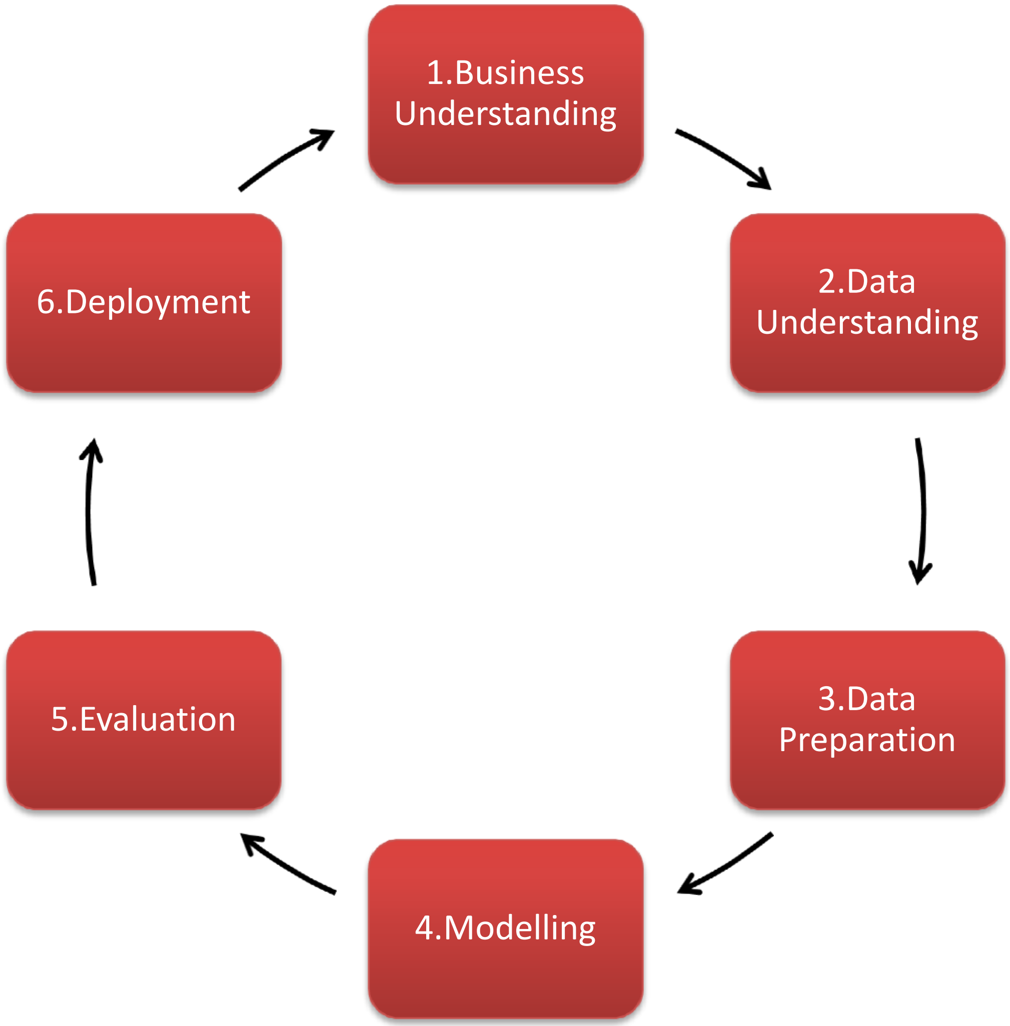
- Premium, accredited partner to IBM & SAS specialising in the SPSS & SAS Advanced Analytics suites.
- Team each has 20+ years of experience working in the predictive analytic space - specifically as senior members of the heritage SPSS team

3 Pillars (high level areas for prediction/data mining)

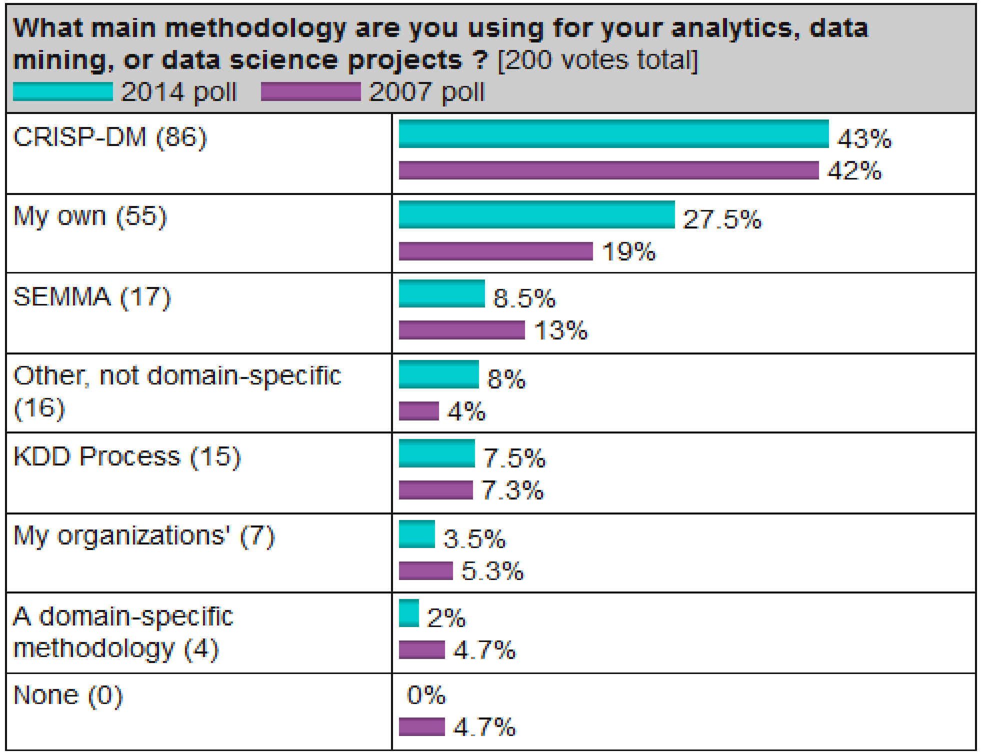
This is the IBM taxonomy which is a good reference for most (if not all) data mining projects

Generally speaking – in each area – we have a process/lifecycle and a series of events that we look to **predict** and **profile**

The CRISP-DM process



KDnuggets 2014 poll

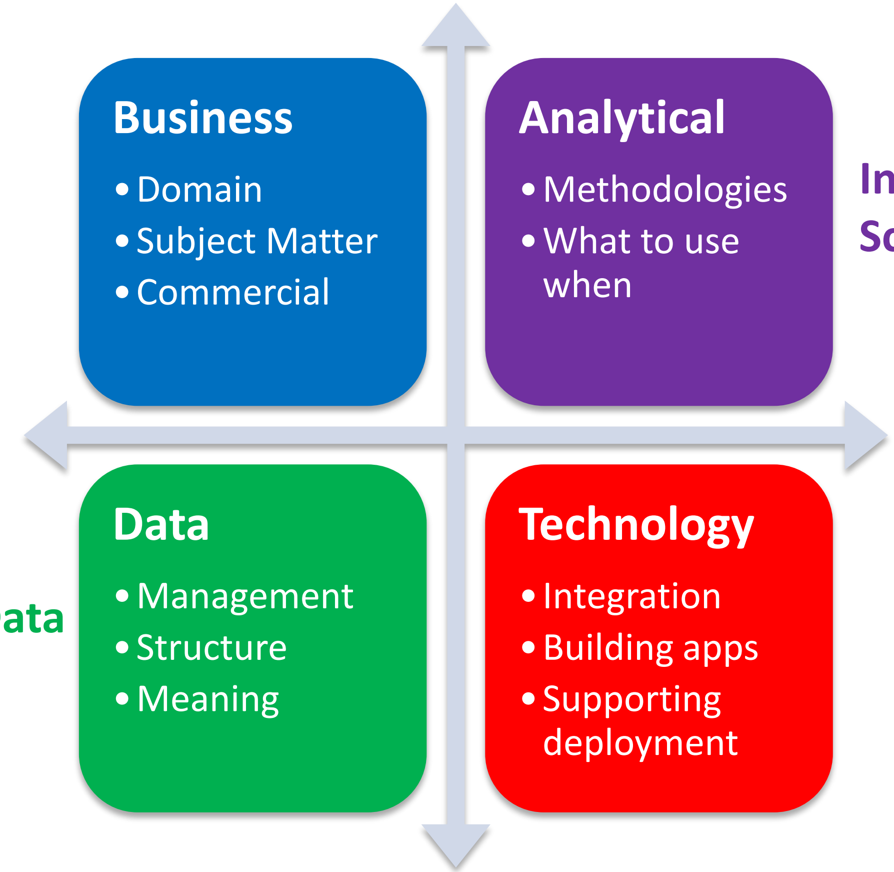


Regional distribution of voters was

- US/Canada, 45.5%
- Europe, 28.5%
- Asia, 14%
- Latin America, 9.5%
- Other, 2.5%

http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

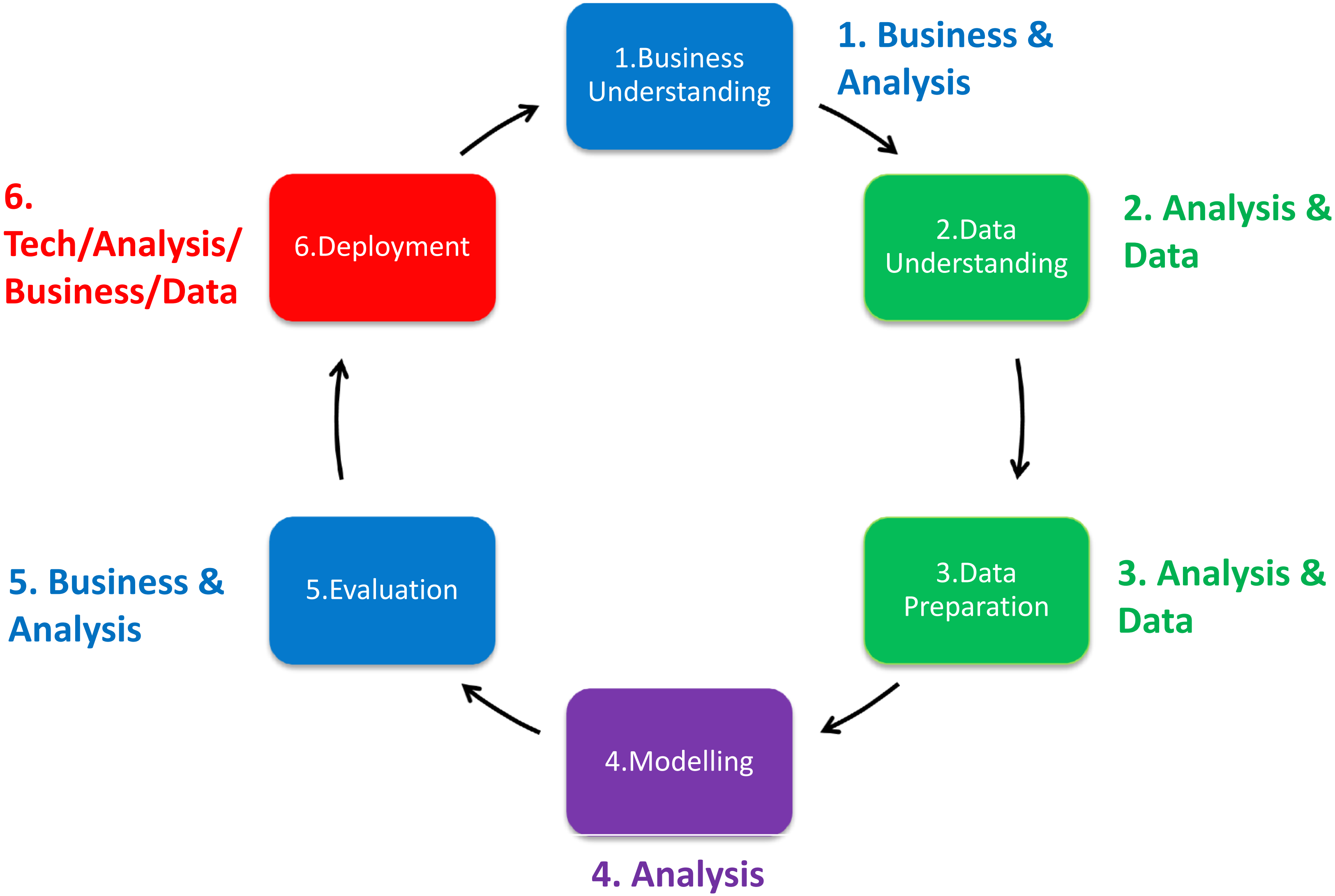
People and Roles



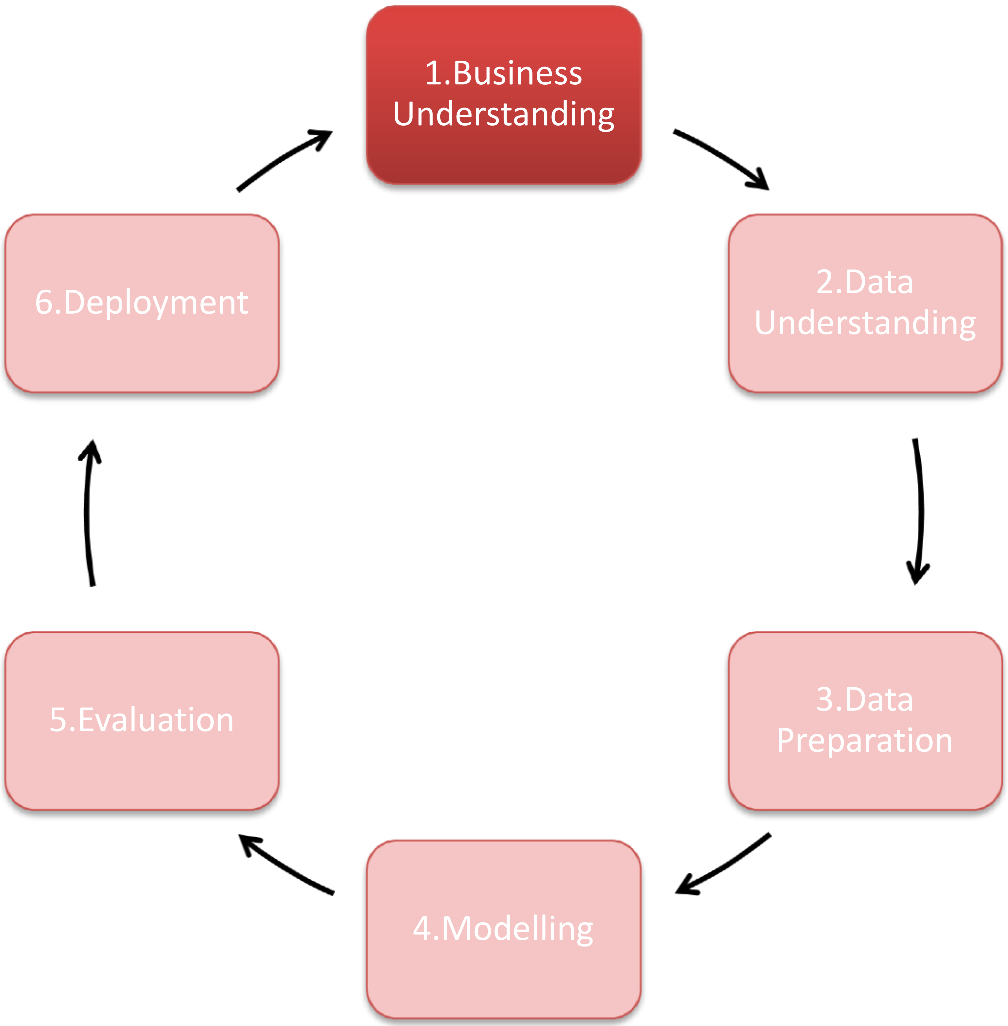
Includes the "Data Scientist" role

Includes the "Data Engineer" role

Roles and Steps in the CRISP process



The CRISP-DM process



1. Business understanding

- Get a clear understanding of the business objectives
  - To reduce churn rates
  - To acquire valuable customers
  - To cross-sell/up-sell
  - To prevent fraud
- Agree success criteria
  - To reduce out annual churn rate from 5% to 3%
- Assess the situation
- Translate to analytical objectives (if possible)
- Evaluate the cost/benefit
- Clearly understand how action can be taken based on the likely outcomes
  - How to deploy
- Document relevant resources, constraints, systems

A selection of example business objectives

- A water company wants to reduce pollution
- An on-line gaming company want to identify fraudulent bets
- A charity wants to increase supporter lifetime value
- A multi-channel subscription-based magazine want to improve renewal rates
- Local government planners want to know how likely a ward is to sustain next year
- A shipping company wants to identify containers that are likely to contain smuggled items
- A coffee retailer wants to understand what effect price changes will have on demand
- A hospital wants to know how many A&E staff to deploy on each shift
- An on-line retailer wants to increase their repurchase rates

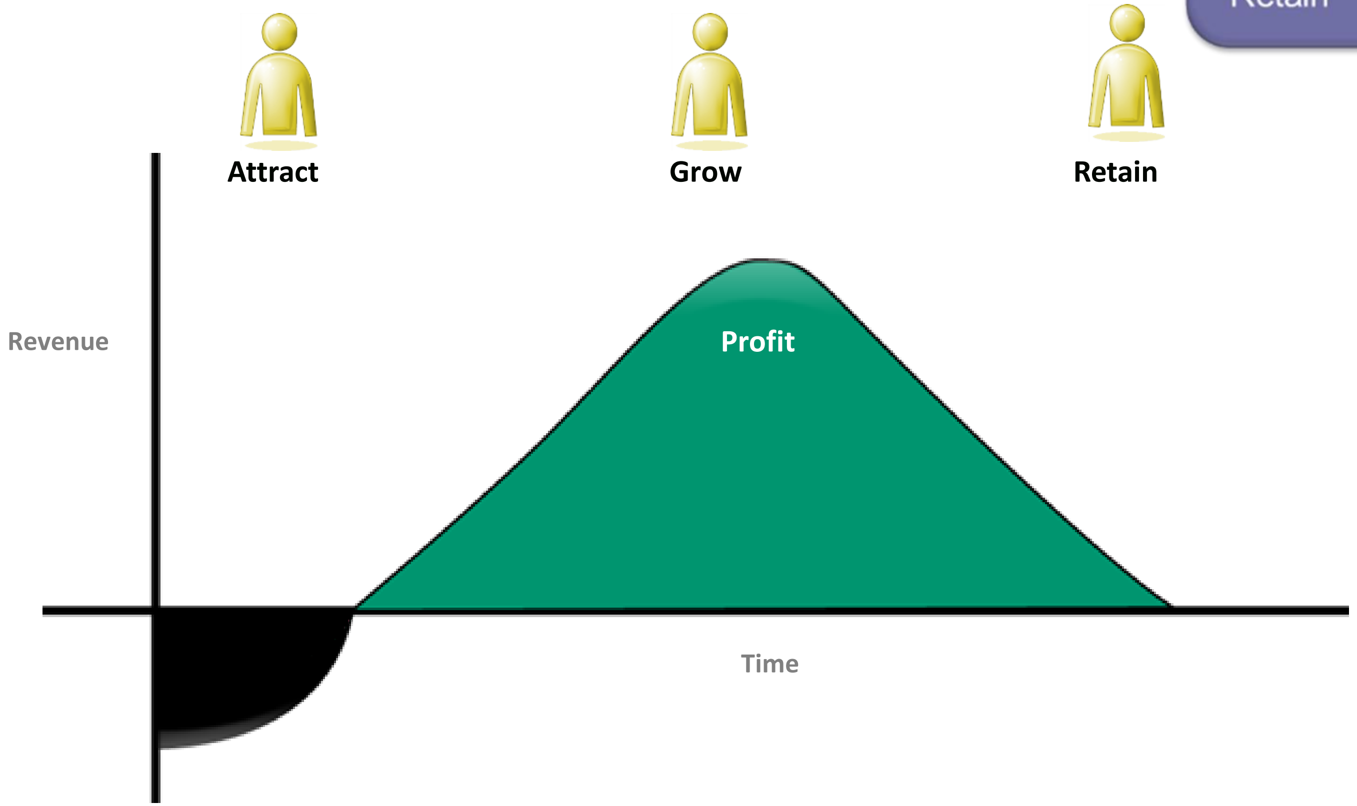
1. Business understanding – Worked Example

- A Retailer wants to
  - Increase revenue and profit
  - By increasing average/total customer Lifetime Value (LTV)
  - They believe they are below their competitive set
- Part of the **strategy** to achieve this is to increase repurchasing
  - They believe they have an issue with repurchase rates for digital customers
- **Success Criteria:**
  - They have a cost constraint – when looking to incentivise repurchase
    - They need to identify at least 40% of customers most likely to repurchase within 20% of all customers
- **Analytically speaking** they want to:
  - **Score** each first purchase customer with a **propensity to repurchase**
  - **Identify** the b of repurchase (or single purchase)

1. Business Understanding - Hypotheses

- The retailer - Business and Analysis teams – develop a set of hypothesised drivers of repurchase. These include:
  - Channel experience
  - Products bought (in first purchase)
  - Value of first purchase
  - Whether the first purchase was a promotion
  - Whether subsequent promotions were made
    - And when (timing)
  - Customer life stage
  - Timing of first purchase; month, day of week, time of day
  - Delivery method
    - Reserve and collect, express, etc.

Predictive Analytics for CRM

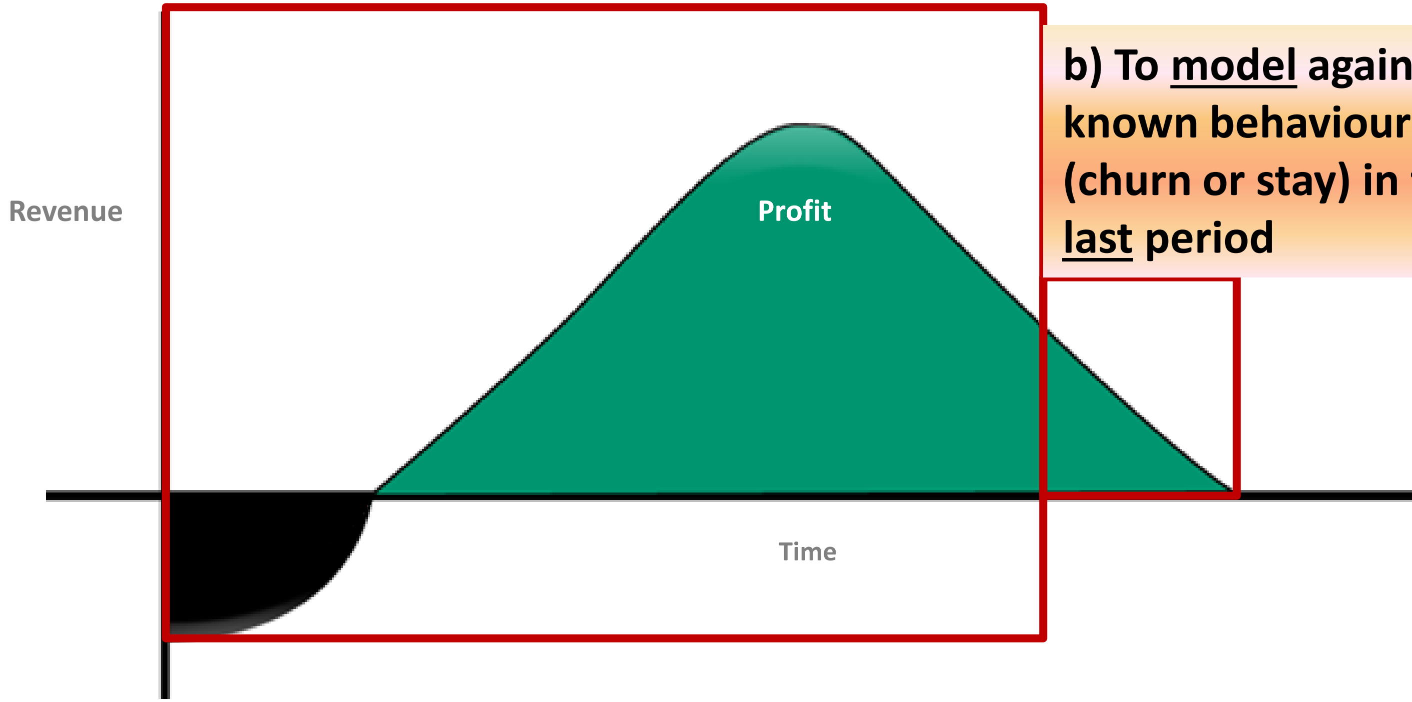


We can model to <u>predict</u> and <u>profile</u> any event across the customer lifecycle…

Modelling Data Window – Churn

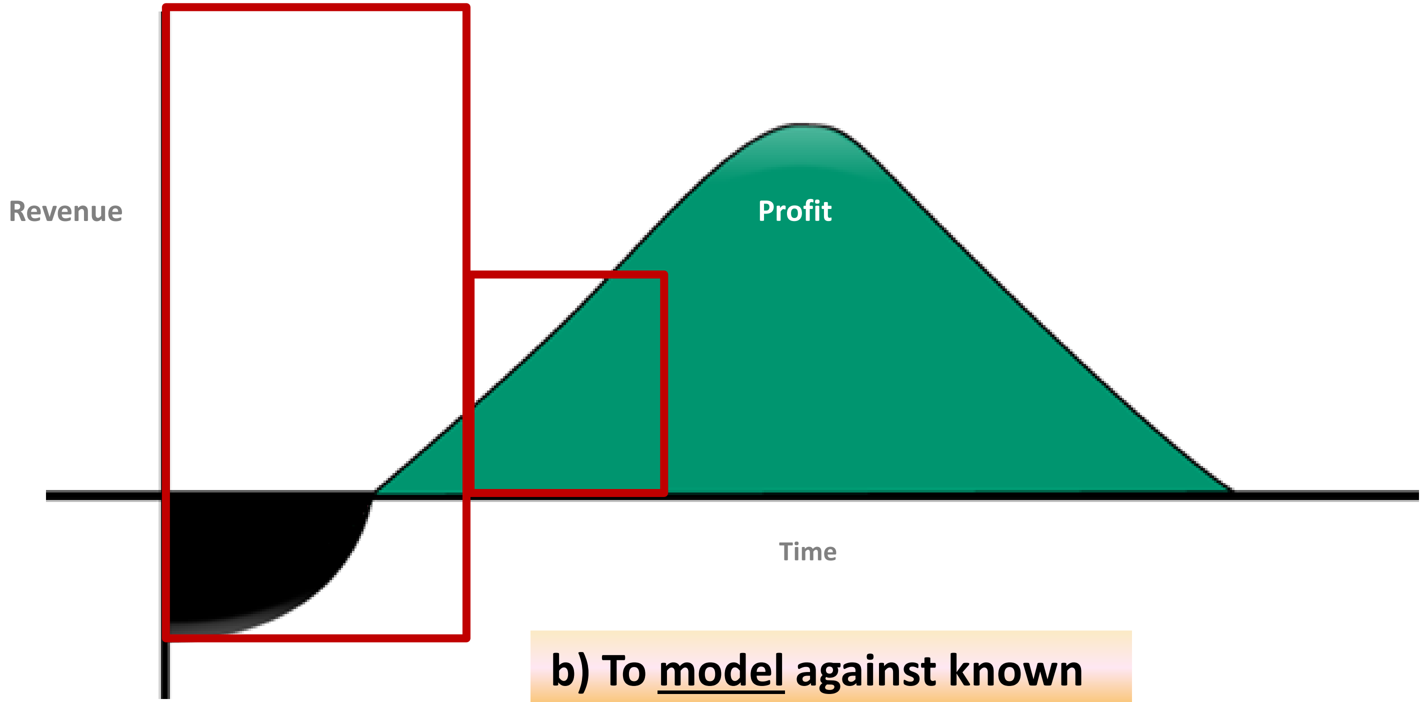
a) Use Data we have on the customer to the <u>time before the last period</u> (e.g. month)

b) To <u>model</u> against known behaviour (churn or stay) in the <u>last</u> period



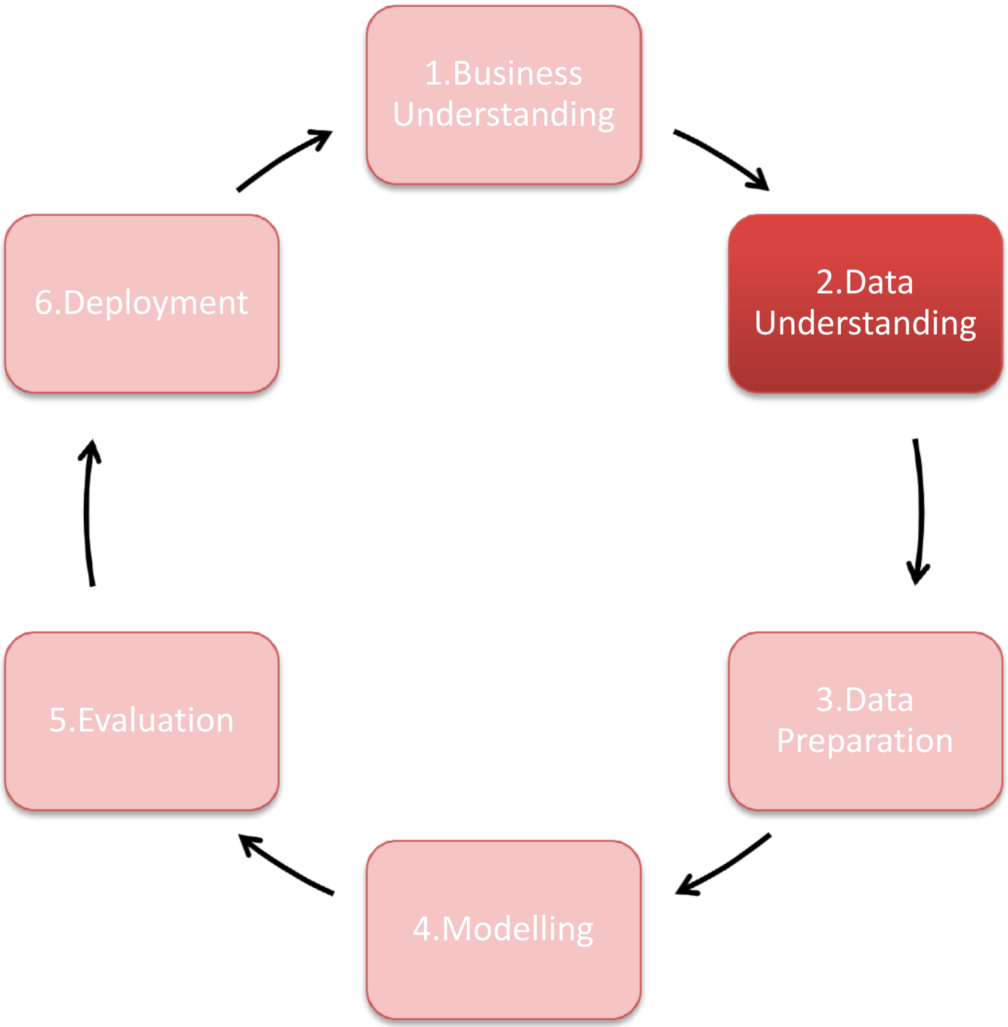
Modelling Data Window – Repurchase

a) Use Data we have on the customer at the <u>point of first purchase</u>



b) To <u>model</u> against known behaviour after that point (repurchase or not)

The CRISP-DM process



2. Data understanding – High Level

- Identify the **data sources** and **fields** which may have a bearing on the business/analytical objectives
- Review data schemas and any other data documentation
- What looks relevant?
- What are the formats?
  - Databases, text files, excel, etc.
- What are the fieldnames?
  - Metadata
- Crucially … what is the likely **target** field that maps to the business objective e.g.
  - Customers purchasing
  - Machinery failing
  - Revenue/Profit/ROI
  - Visits to the web site
  - Denial of service attacks detected
  - Customers churning

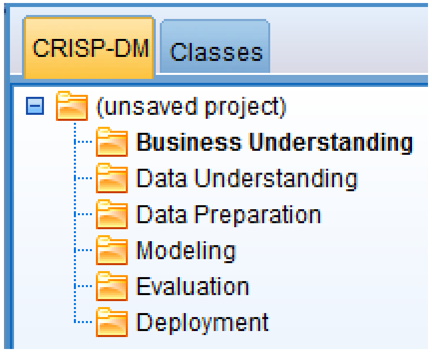
2. Data understanding – Low level

- **Explore** the data
- Typically looking for **patterns** between fields
- Using uni- and bi-variate analyses
  - Examine fields one-by-one or in pairs
  - Often using visualisation tools
- Test hypotheses
  - E.g. Age of donor is a predictor of value
- Validate data
  - Identifies any issues involving anomalies
- Develops understanding and informs modelling

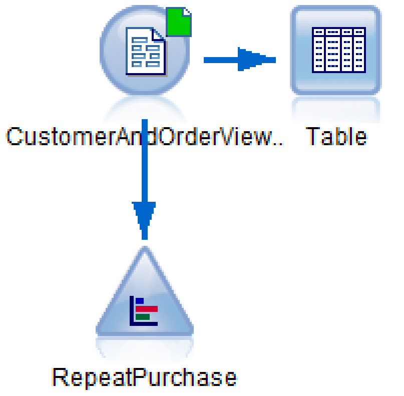
2. Data understanding – Worked Example

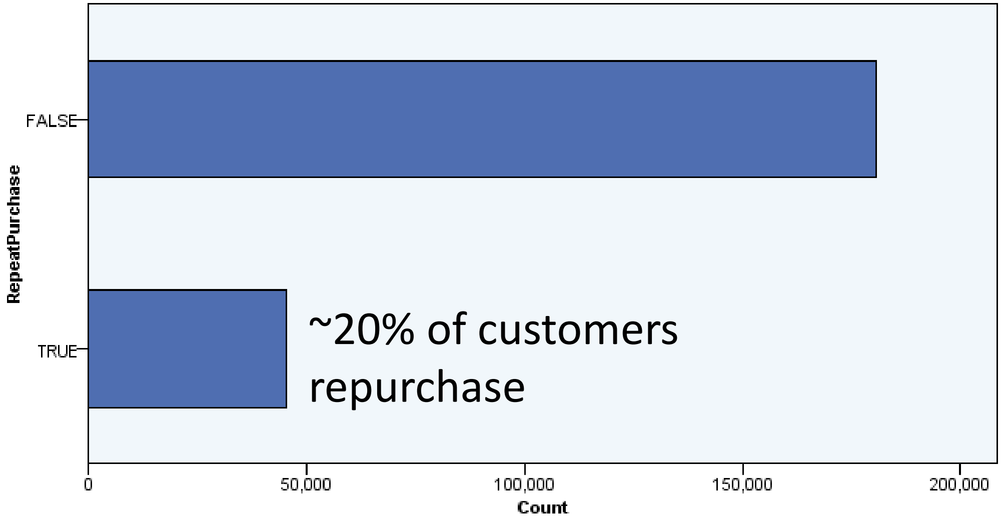
We will use **IBM/SPSS Modeler** to demonstrate the software-related steps in the CRISP process

Modeler maps to that process



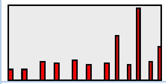
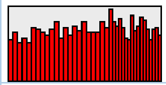
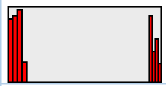
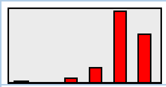
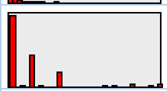
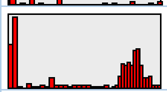
We can build a Modeler "**stream**" to map to that process. Starting with our **Target** field





Base: 225,793 customers

2. Data understanding – Audit to Validate

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev
FirstOrderMonth		Continuous	1	12	7.837	3.101
Sequence		Continuous	2	276158	140569.181	79657.985
URN_Customer		Continuous	151914482...	31737012258	22090012826.481	7351908174.182
FirstOrder_Channel		Categorical	--	--	--	--
FirstOrder_OrderValue		Continuous	0.000	5984.300	91.408	117.850
FirstOrder_Delivery_Type		Categorical	--	--	--	--
FirstOrder_odate		Continuous	2006-08-11	2011-12-05	--	--
FirstOrder_NumberOfOrderItems		Continuous	1	83	2.340	2.515

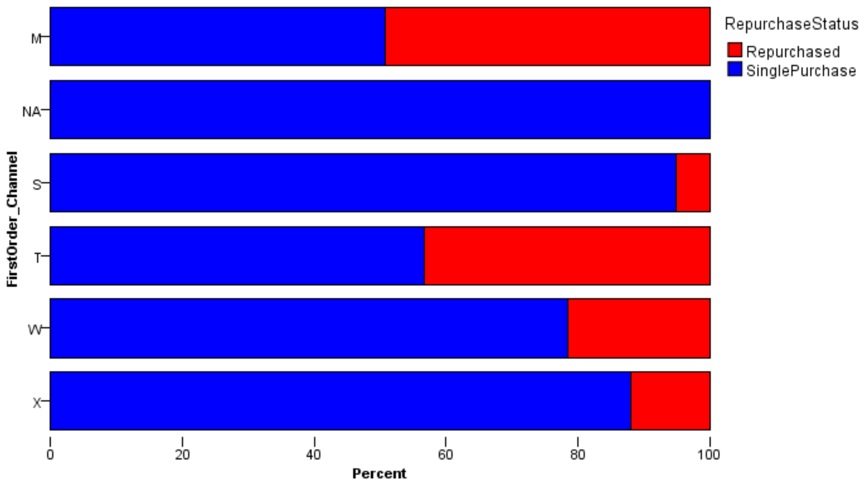
It is important to check the fields we will use for modelling particularly looking for

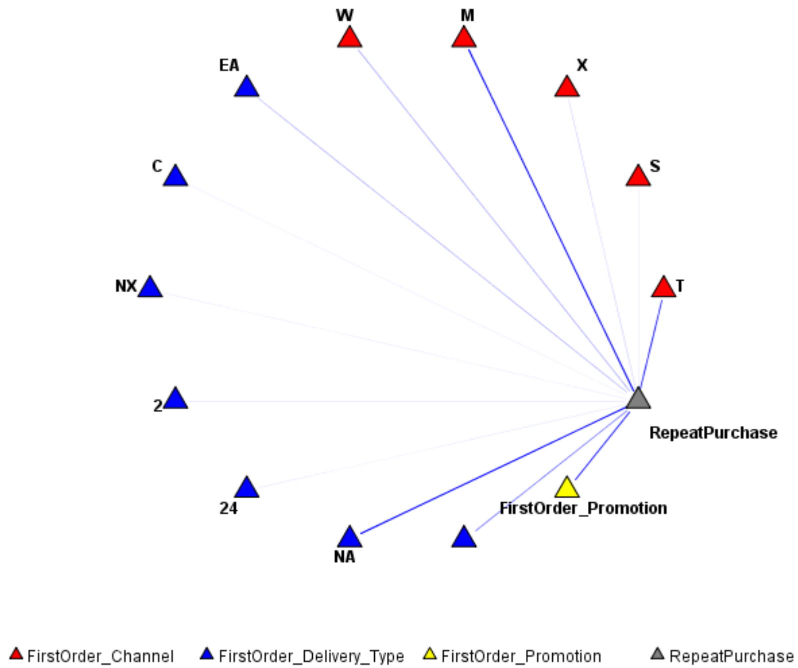
- Missing values
- Unusual values
- Invalid values

Base: 225,793 customers

2. Data understanding – Visually exploring relationships

A stacked bar confirms our hypotheses that some channels are more likely to lead to (/drive) repurchase

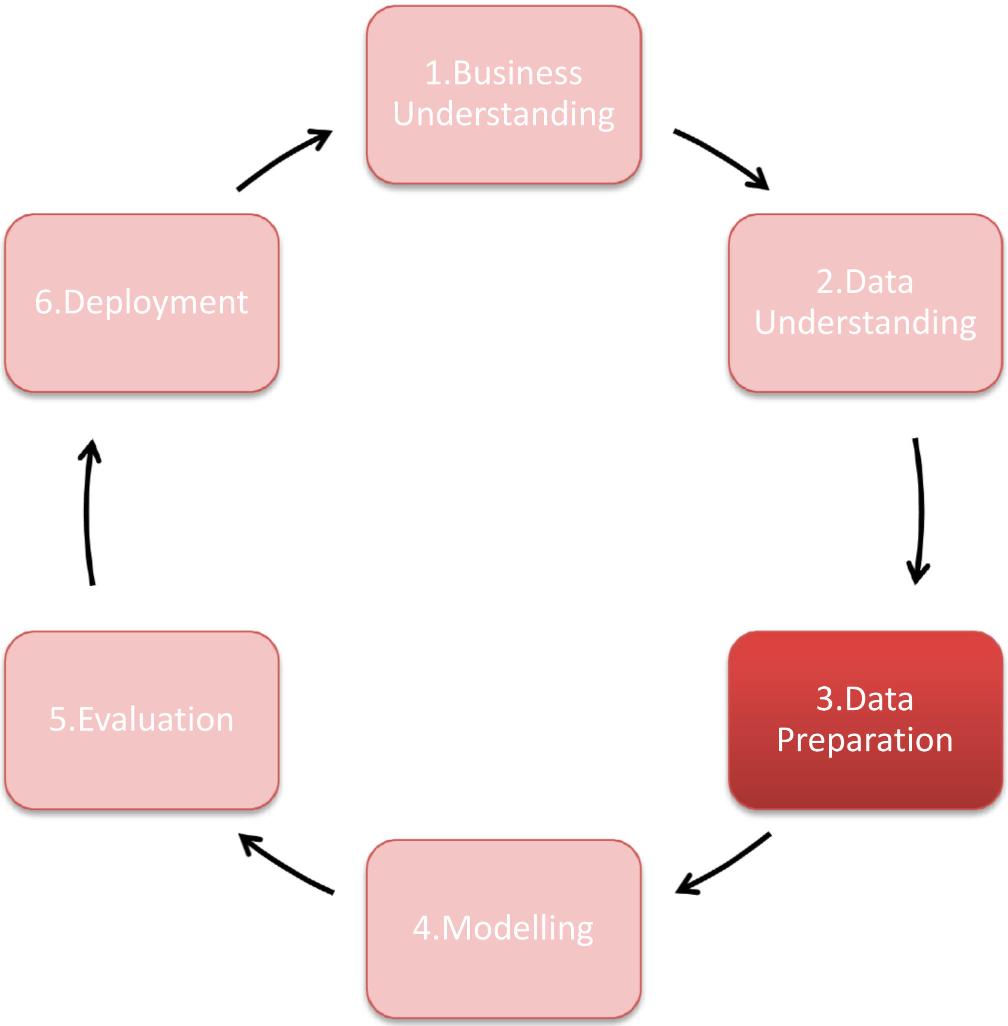




A web plot also shows that some delivery types and promotions also lead to a higher repurchase %

Base: 225,793 customers

The CRISP-DM process

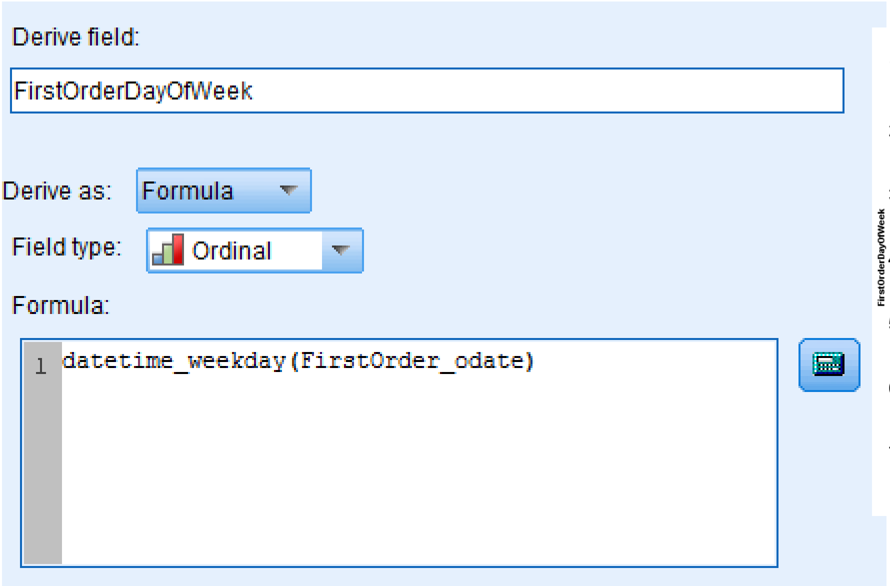


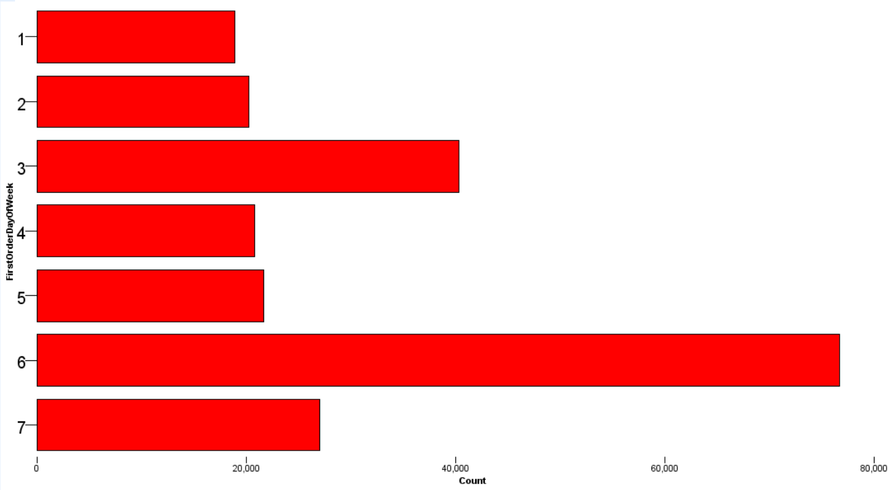
3. Data Preparation

- Data Understanding helps design this step
- Together with Data Understanding this can be more time consuming than expected
  - Sometimes 80% of a project
  - Especially for newer projects
- Typically integrates data from different sources
- Aggregates data
- Create composite measures
  - E.g. band variables
  - Apply formulae e.g. compute annualised figures and other ratios
- Comparable to ETL (Extract Transform Load)

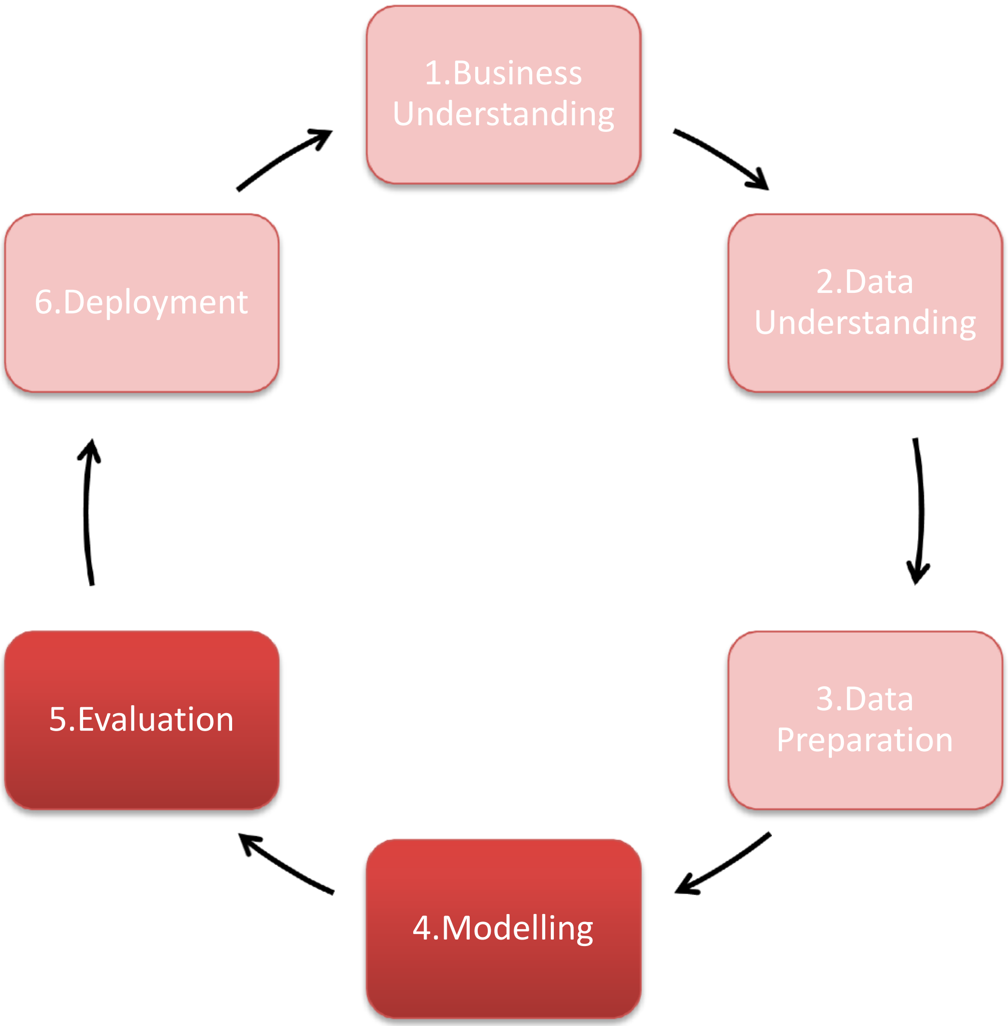
3. Data Preparation – Repurchase example

- In our example we received an extract from the Customer Data Warehouse in a single file
  - Often we can liaise with the DBA and the DBA will do some or all of the data prep
- Most often we need to integrate data from multiple files / database tables
- We didn't get all the fields we hypothesised about but we can create one of the missing ones (day of week) using a **Derive** node





The CRISP-DM process



4.Modelling

- Apply a variety of modelling techniques
- Candidate list identified during understanding phase
  - Driven by data types (see later)
  - Constrained by available tools
- 2 broad styles:

  a) **Hypothesis led**. Add the fields/predictors that we believe are driving the outcome

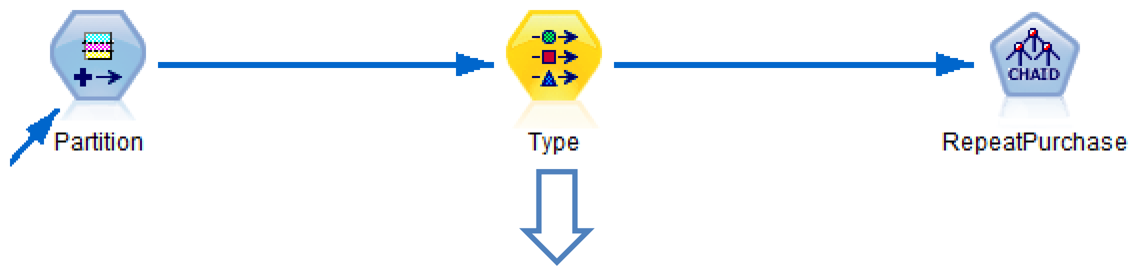
  b) **Data led**. Add more fields at the beginning and incrementally reduce (and/or let the algorithms do that)
- The best performing modelling algorithm is a function of the specific data/problem

4.Modelling – Repurchase rates

The **Partition** node creates the 70:30 split

The **Type** node is where we choose **Inputs** (predictors/drivers) and the **Target**

The modelling node (in this case we chose a **CHAID** node) builds the model

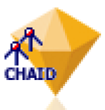


Field	Measurement	Values	...	Check	Role
Sequence	Continuous	[2,27...		None	None
URN_Customer	Continuous	[151...		None	None
FirstOrder_NumberOfOrderItems	Continuous	[1,83]		None	Input
FirstOrder_Channel	Nominal	M,NA...		None	Input
FirstOrder_OrderValue	Continuous	[0.0,5...		None	Input
FirstOrder_Delivery_Type	Nominal	"","2",...		None	Input
FirstOrder_odate	Continuous	[200...		None	Input
FirstOrderMonth	Continuous	[1,12]		None	Input
FirstOrder_Promotion	Flag	T/F		None	Input
RepeatPurchase	Flag	TRU...		None	Target
Title	Nominal	"",Dr,...		None	Input
Mailable	Nominal	"","D...		None	None
Emailable	Nominal	"","D...		None	None
goneaway	Flag	"Y ...		None	None
RepurchaseStatus	Flag	Singl...		None	None
FirstOrderDayOfWeek	Ordinal	<Rea...		None	Input
Partition	Nominal	"1_Tr...		None	Partition

5.Evaluation

- Essential that the models are tested against unseen data
- Typically the data is partitioned into 2 (or 3) sets at random e.g. 70%:30%
  1. Training (modelling) set
  2. Test (holdout) set
  3. Evaluation set
- Evaluate against the **success criteria** agreed in the understanding phase
- Often it is about how well the model performs against a given value criteria e.g. revenue
  - Defined in **Data Understanding** phase

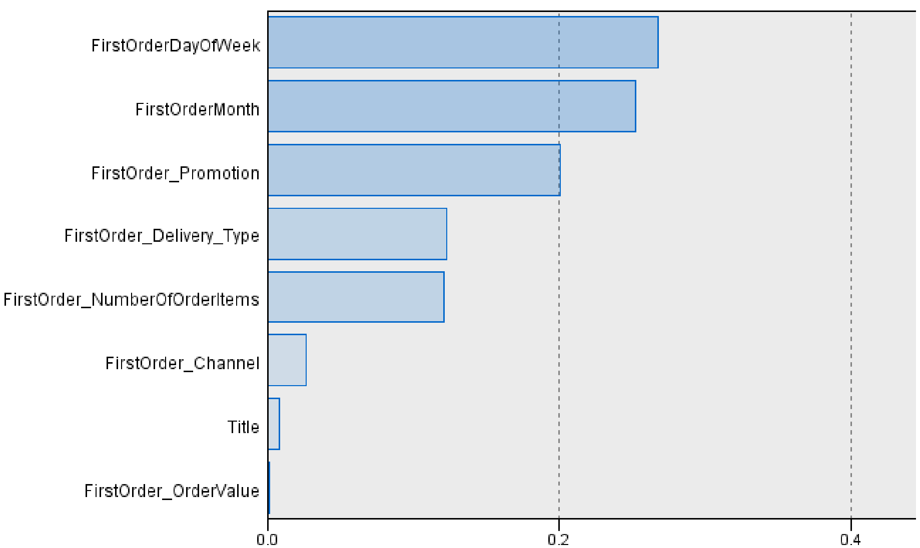
5. Evaluation – Worked Example



RepeatPurchase

The “nugget” is the built model.
When we edit it we see more detail of the model including:

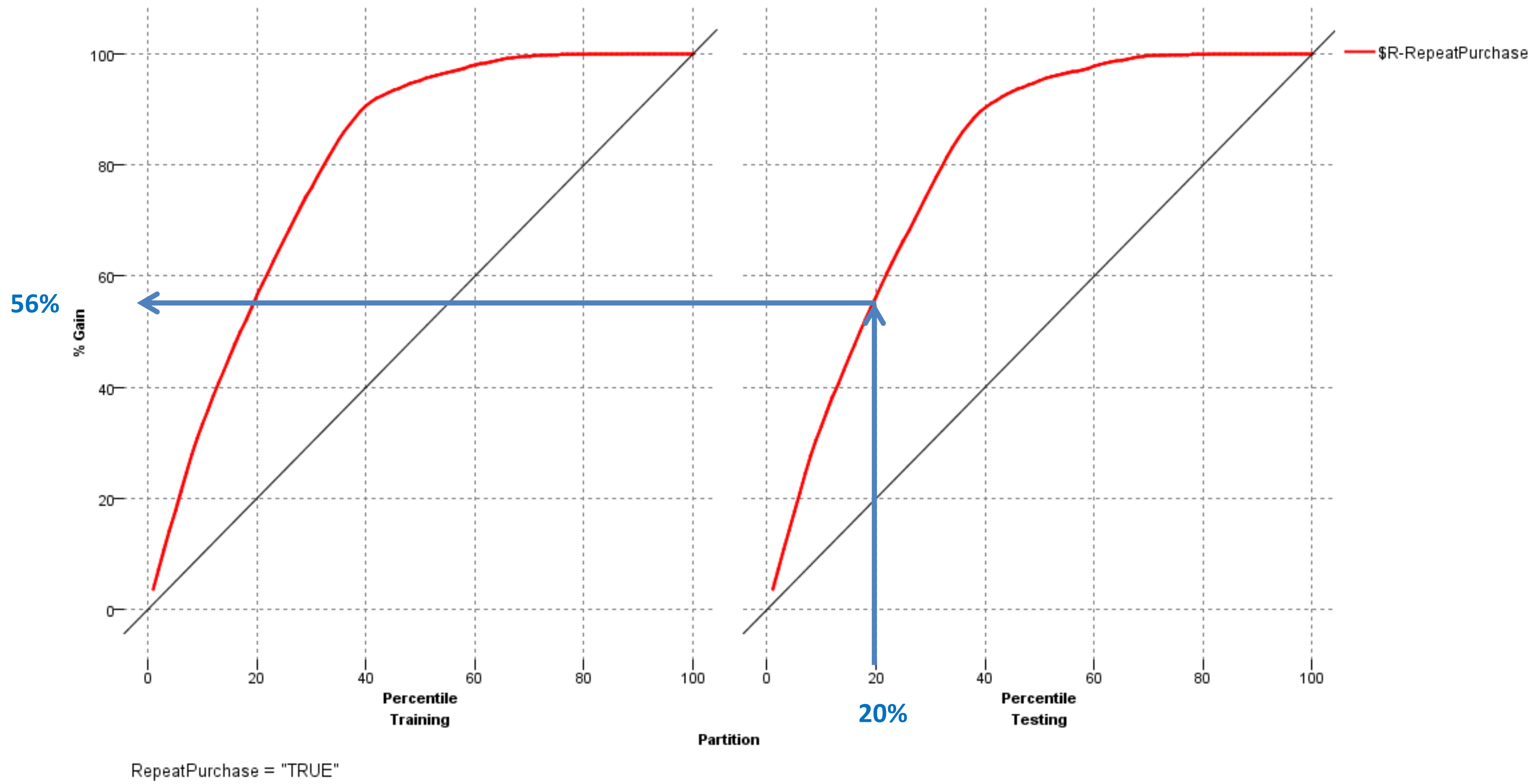
Predictor Importance



Rules/repurchase profiles

Rule 1 for TRUE (6,121; 0.704)
if FirstOrderMonth > 9
and FirstOrderMonth <= 10
and FirstOrderDayOfWeek > 5
and FirstOrder_Promotion in ["F"]
and Title in ["Miss" "Dr" "Mrs" "Ms"]
and FirstOrder_Delivery_Type in [""]
then TRUE

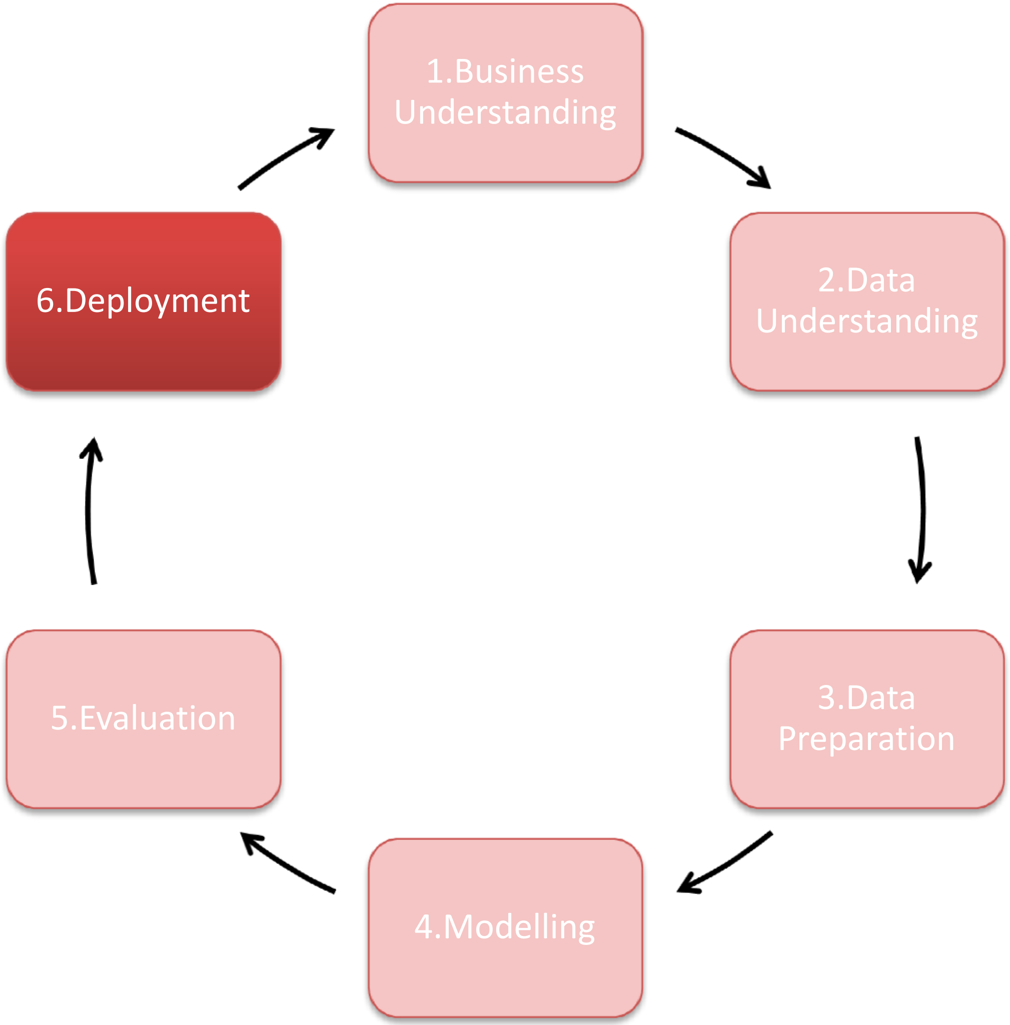
5. Evaluation – Predictive Accuracy



Recall our success criteria (in Business Understanding) was to be able to predict 40% of re-purchasers among 20% of customers

This model beats that as it gains (predicts) **56% of re-purchasers in the top 20%**

The CRISP-DM process

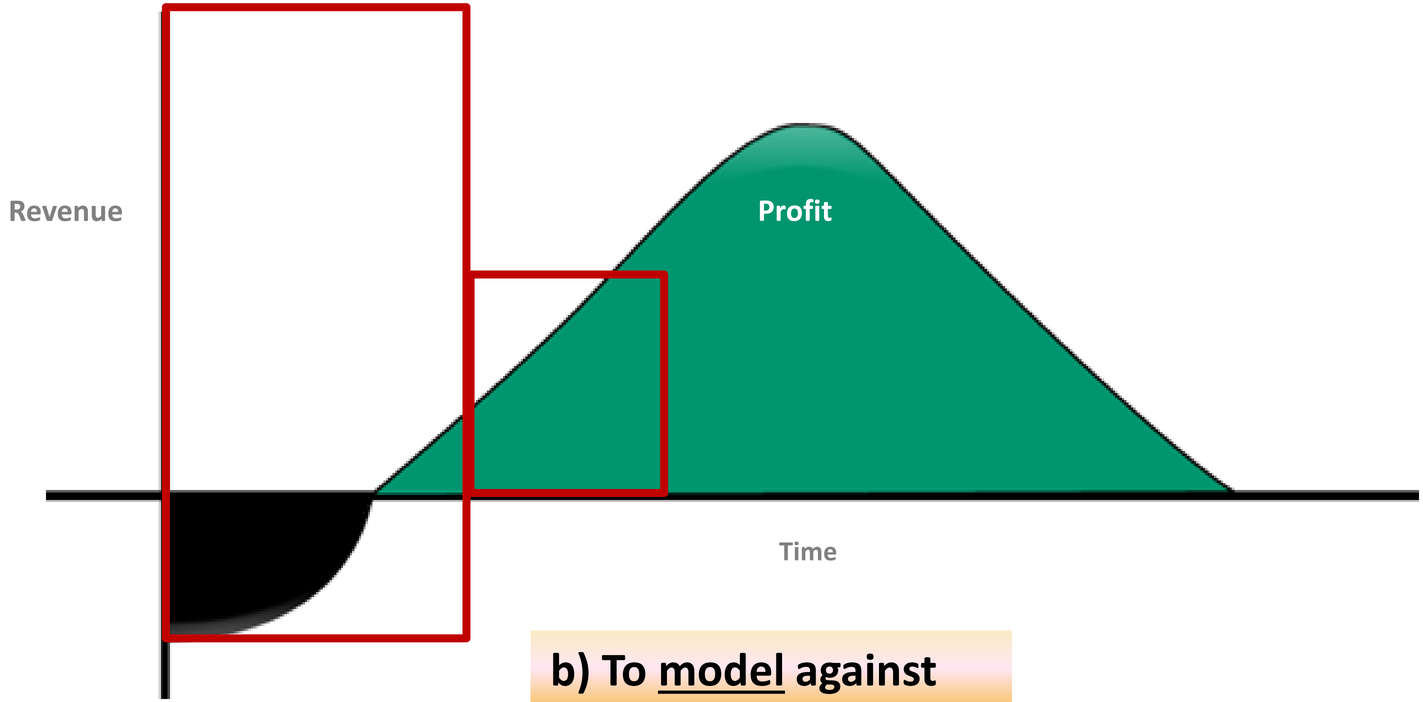


6.Deployment (1)

- Could be as simple as a list of names and predictions/scores
  - E.g. a mailing list
- Could be as complex as a model encapsulated as a computer program and embedded in an operational system to predict in real time and automate decisions
  - E.g. a model embedded in a system which sends alerts and triggers
- Could be embedded in a **What-if?** simulator
- Important to distinguish between a model in the modelling and deployment phases
- Typically…
  - In the modelling phase many different models and modelling options are built and evaluated
  - In the deployment phase the winning model(s) are fixed
    - E.g. we deploy a decision tree with a fixed shape

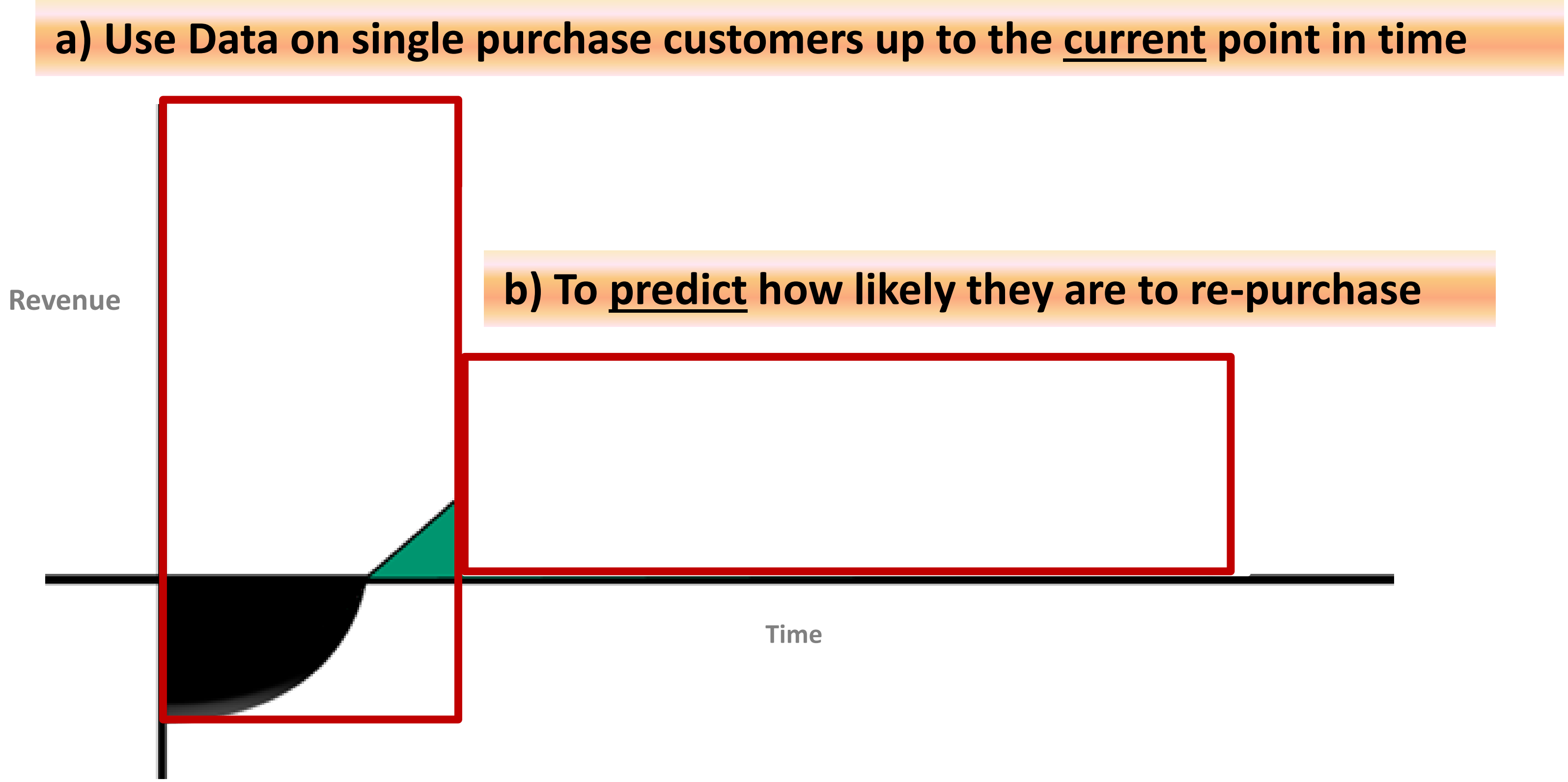
Modelling Data Window – Repurchase

a) Use Data we have on the customer to the <u>time before the last period</u> (e.g. month)

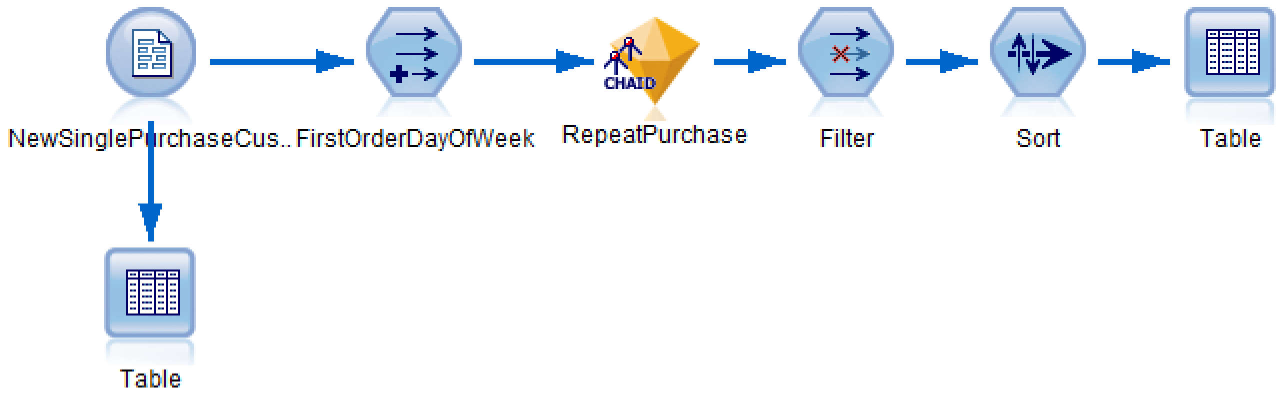


b) To <u>model</u> against known behaviour (repurchase or not)

Modelling Data Window – Scoring new customers



6. Deployment – Repurchase example



In one form or other – and there are many ways we can do this – we plug the winning model into a deployment process. Here as another modeller stream

The predictions are often sent to a database or file …

or directly into an operational system

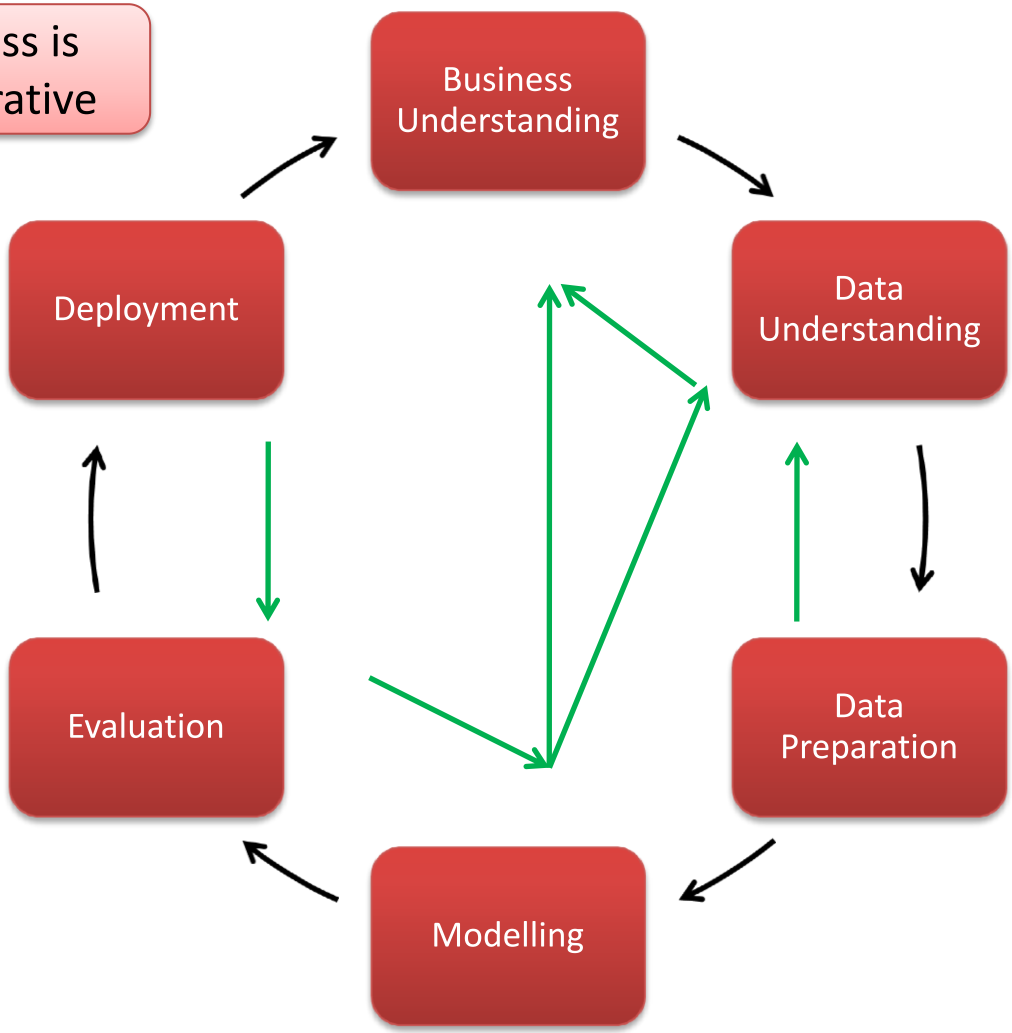
	URN_Customer	PropensityToRepurchase
1	16111698240	0.733
2	31191906246	0.733
3	31629870246	0.733
4	31226022246	0.733
5	31221750240	0.733
6	30504414240	0.733
7	30477786246	0.733
8	30282336240	0.733
9	30249996288	0.733
10	30465336240	0.733
11	31214964240	0.733

6.Deployment (2) (Monitoring)

- If we did our job properly then the deployed model should correspond to what we saw in evaluation
  - Other factors may intervene
- Ongoing evaluation ("monitoring") still needs to happen if models are to be used over time
  - Some models have a longer shelf life than others
- More recently there has been some development of models which adapt/correct themselves to changing circumstances
  - Some level of re-modelling to improve accuracy
    - "Self adapting"
  - More commonly this is achieved through the concept of **champion/challenger** modelling or **model refresh** approaches

The CRISP-DM process reprised

The process is usually iterative



Executing a predictive project - summary

- A Predictive Analytics project can be more like a **Research & Development** project
  - Can we build a successful model?
  - Has anyone done this before?
  - What is the **risk** that we cannot achieve the objectives?
- Hence projects can fail
- It isn't just about the analyst
  - Larger projects usually need a larger (multi-disciplined) team
- Each stage in the CRISP-DM process has detailed tasks and outputs
  - More detail at:

  http://www.sv-europe.com/crisp-dm-methodology/

Working with Smart Vision Europe Ltd

- As a premier partner we resell software to you directly
  - We're agile, responsive and generally easier to deal with
- As experts in SPSS / SAS / Analytics / Predictive Analytics we will
  - deliver classroom training courses
  - offer side by side training support
  - offer "skills transfer" consulting
  - run booster and refresher sessions to get more from your SPSS licences
  - Give no strings attached advice
- We're approachable, friendly and pride ourselves on our no-nonsense approach
  - If you have questions please do email us info@sv-europe.com
  - If you'd like advice, do get in touch and we'll help if we can!

Contact us:

+44 (0)207 786 3568
info@sv-europe.com
Twitter: @sveurope
Follow us on Linked In
Sign up for our Newsletter

Thank you