Planning Successful Predictive Analytics Projects

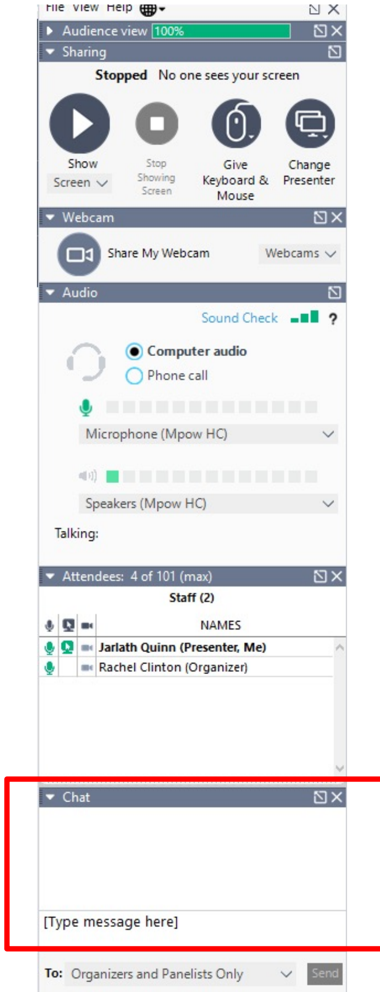
Jarlath Quinn

www.sv-europe.com

A SELECT INTERNATIONAL COMPANY

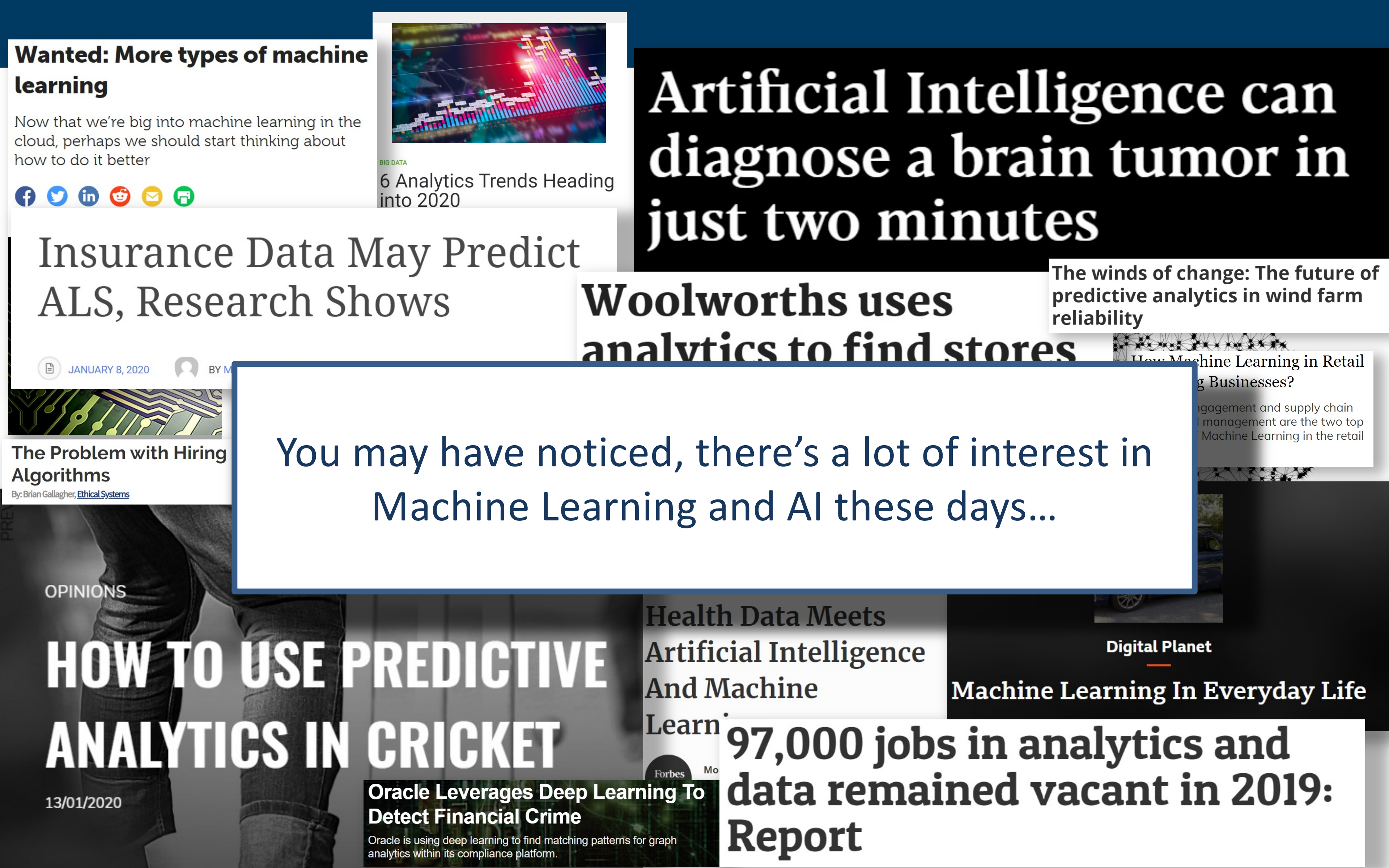
FAQ's

- Is this session being recorded? Yes
- Can I get a copy of the slides? Yes, we'll email a PDF copy to you after the session has ended.
- 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.



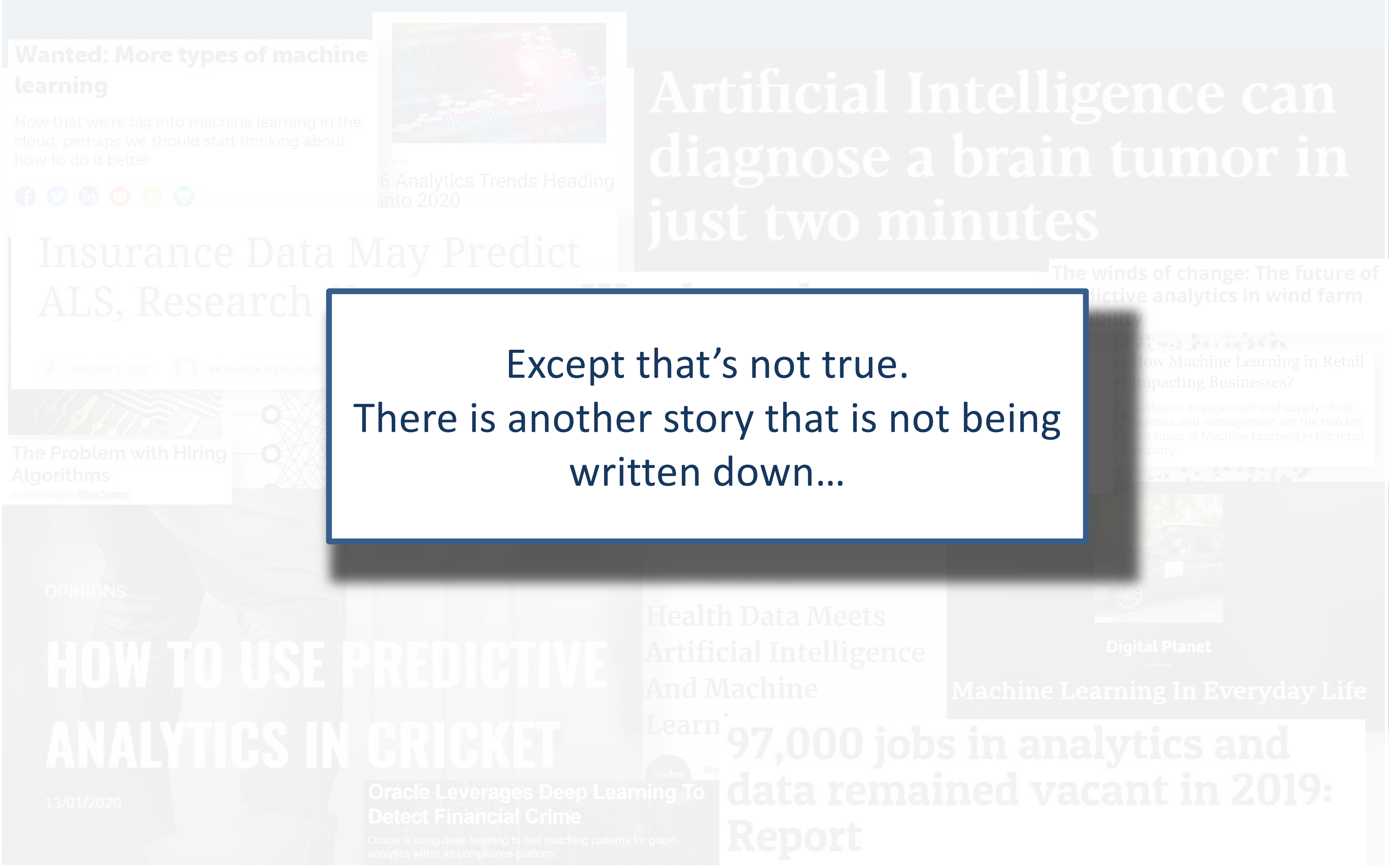
- Premier accredited partner to IBM and Predictive Solutions specialising in advanced analytics & big data technologies
- Work with open source technologies (R, Python, Spark etc.)
- Team each has 15 to 30 years of experience working in the advanced and predictive analytics industry

- Deep experience of applied advanced analytics applications across sectors
  - Retail
  - Gaming
  - Utilities
  - Insurance
  - Telecommunications
  - Media
  - FMCG


Wanted: More types of machine learning
Now that we're big into machine learning in the cloud, perhaps we should start thinking about how to do it better
BIG DATA
6 Analytics Trends Heading into 2020
Artificial Intelligence can diagnose a brain tumor in just two minutes
Insurance Data May Predict ALS, Research Shows
JANUARY 8, 2020
Woolworths uses
The winds of change: The future of predictive analytics in wind farm reliability
You may have noticed, there's a lot of interest in Machine Learning and AI these days…
The Problem with Hiring Algorithms
By: Brian Gallagher, Ethical Systems
OPINIONS
HOW TO USE PREDICTIVE ANALYTICS IN CRICKET
13/01/2020
Health Data Meets Artificial Intelligence And Machine
Digital Planet
Machine Learning In Everyday Life
97,000 jobs in analytics and data remained vacant in 2019: Report
Oracle Leverages Deep Learning To Detect Financial Crime
Oracle is using deep learning to find matching patterns for graph analytics within its compliance platform.

You might get the impression that every organisation is rapidly and successfully employing these approaches

Except that’s not true.
There is another story that is not being written down…



Most organisations are doing very little with advanced analytics.

Even when they have embarked on a project, these initiatives are often fragmented across the business, highly-experimental in nature, vaguely defined and prone to making little tangible difference to the real world.

In other words, many of these initiatives are effectively failures.

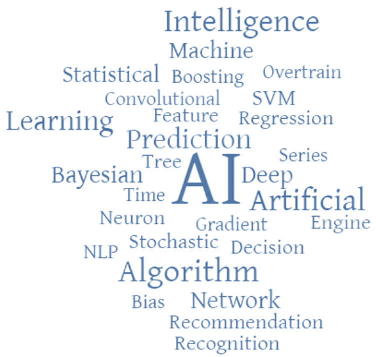
In collaboration with

- A 2020 report by **MIT Sloan** in collaboration with **BCG** confirmed what many industry insiders have known for years:
  - *only a small minority of companies manage to make their initial AI projects succeed*.
- Based on global survey of more than 3,000 managers and scholars in 29 industries, the authors discovered that ***a mere 11%*** of organisations saw significant financial benefits from their AI programmes.
- The report states AI challenges are not solved by "having the right data, technology, and talent, organized around a corporate strategy"
- Rather they require "large-scale organizational shifts in mindsets".

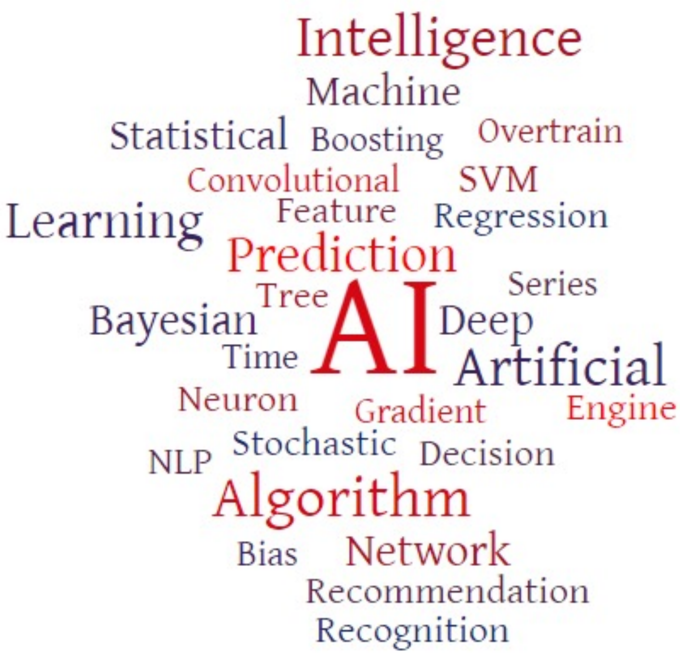
In collaboration with

- A [2020 report](#) by **MIT Sloan** in collaboration with **BCG** confirmed what many industry insiders have known for years:
  - … *projects*
- Bas… 9 industries, the… t financial be…
- The… , tec…
- Rather they require "large-scale organizational shifts in mindsets".

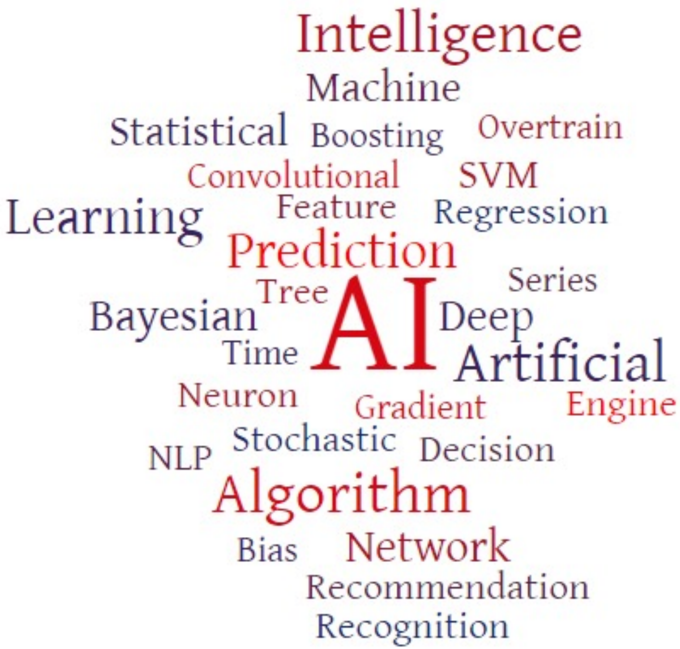
So what did the successful organisations do right?



How Predictive Analytics works in the real world



What we talk about when we talk about AI and Machine Learning



= **Predictive Analytics**

"Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events."

"Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events."

What do these approaches all have in common?

Table (34 fields, 7,043 records)

File Edit Generate

Table Annotations

	h	$R-Churn	$RC-Churn	$R1-Churn	$RC1-Churn	$RI1-Churn	$L-Churn	$LC-Churn
1	ng	No	0.889	No	0.930	11	No	0.734
2	ing	Yes	0.889	Yes	0.705	39	Yes	0.593
3	ng	Yes	0.556	No	0.729	25	No	0.545
4	ing	No	0.778	No	0.812	23	No	0.681
5	ng	No	0.889	No	0.883	8	No	0.677
6	ing	No	1.000	No	0.986	10	No	0.843
7	ing	No	0.889	No	0.883	8	No	0.578
8	ng	Yes	0.556	Yes	0.605	41	Yes	0.505
9	ing	No	0.556	No	0.796	34	No	0.687
10	ng	No	0.778	No	0.779	43	No	0.604
11	ng	Yes	1.000	Yes	0.705	39	Yes	0.643
12	ing	No	0.778	No	0.883	8	No	0.769
13	ing	Yes	1.000	No	0.558	35	No	0.615
14	ing	No	1.000	No	0.986	10	No	0.815
15	ng	No	0.778	No	0.880	26	No	0.604
16	ng	No	0.556	No	0.729	25	No	0.554
17	ng	Yes	0.556	Yes	0.503	29	No	0.511
18	ing	No	0.778	No	0.750	45	No	0.651
19	ing	Yes	0.556	No	0.796	34	No	0.533
20	ing	Yes	0.556	No	0.812	23	No	0.638
21	ng	No	0.778	No	0.883	8	No	0.573

OK

They generate new data

What do we mean when we talk about 'Predictive Analytics'?

- Ironically, it's not *always* about prediction *per se*
- But Predictive Analytics can always *create new data*
- These data take the form of estimates, probabilities, forecasts, recommendations, propensity scores, classifications or likelihood values
- The acid test of an analytical *model* is how accurate these new data are
- But the usefulness of an analytical *application* depends on the *decisions* we take as a result of these new data

How is Predictive Analytics applied?

- **Retail**
  - Promotions, Basket Analysis, Store Clustering, Forecasting
- **Financial**
  - Credit Scoring, Fraud, Mortgage Retention, Product Cross-sell
- **Communications and Media**
  - Retaining Subscribers, Feedback Mining, Content Recommendation
- **Insurance**
  - Satisfaction Modelling, Retention, Fraud, Claim Risk
- **Utilities**
  - Customer Profitability, Problem Resolution, Predictive Maintenance

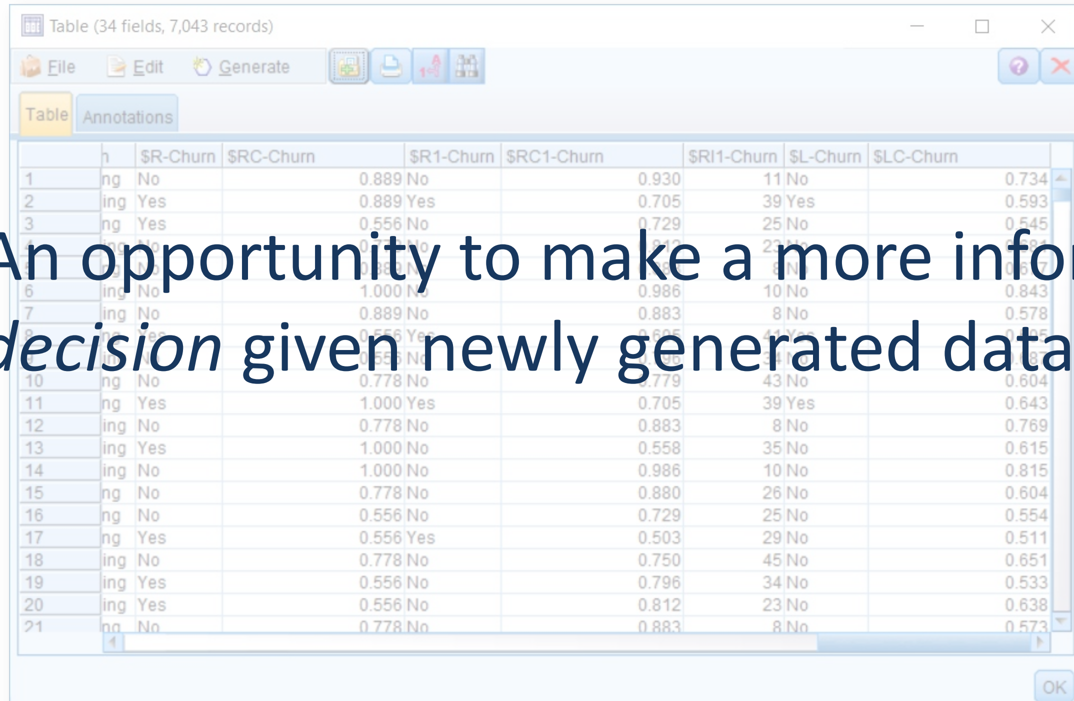
How is Predictive Analytics applied?

- **Tax**
  - Non-compliance, Fraud, Service Quality
- **Charities**
  - Campaign Response, Supporter Segmentation, Legacy Giving
- **Education**
  - Retention, Acquisition, Student Performance
- **Healthcare**
  - Patient Readmission, Patient Safety, Delay Analysis
- **Police**
  - Crime Prediction, Satisfaction Modelling, Outcome Modelling

Predictive Analytics is not...

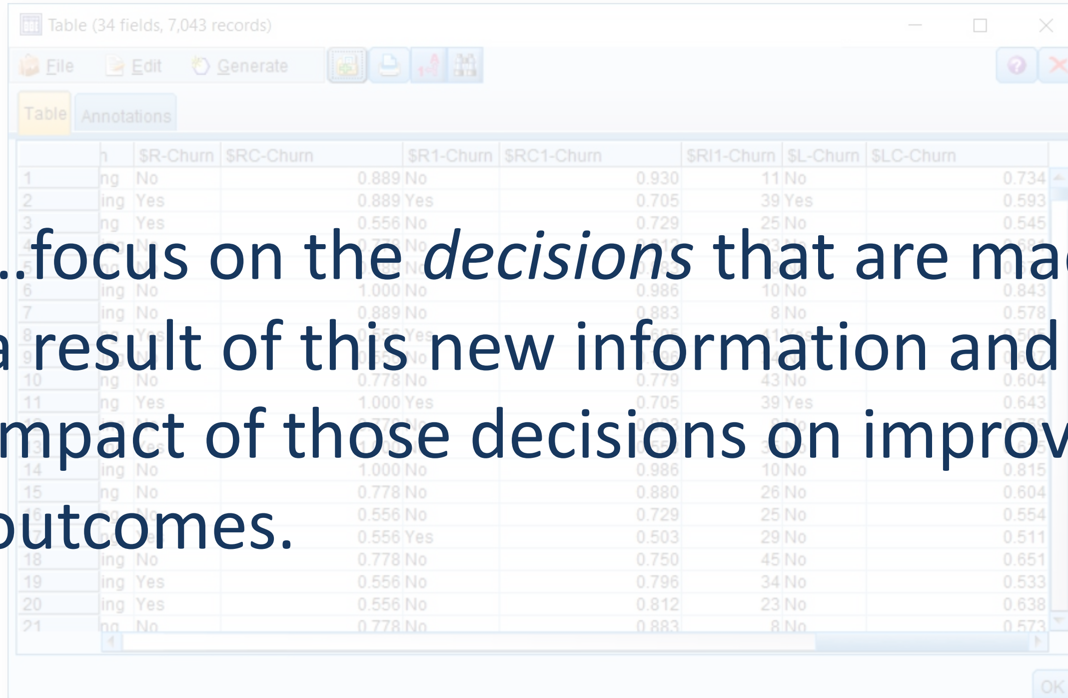
- A super-charged version of BI that is designed to reveal hidden secrets
- An insight platform that will tell you “what to do next…now”
- An approach for calculating the optimal outcomes
- A data visualisation discipline

Predictive Analytics is…



- An opportunity to make a more informed *decision* given newly generated data…

Successful Predictive Analytics initiatives…



- …focus on the *decisions* that are made as a result of this new information and the impact of those decisions on improving outcomes.

What is the right decision if…

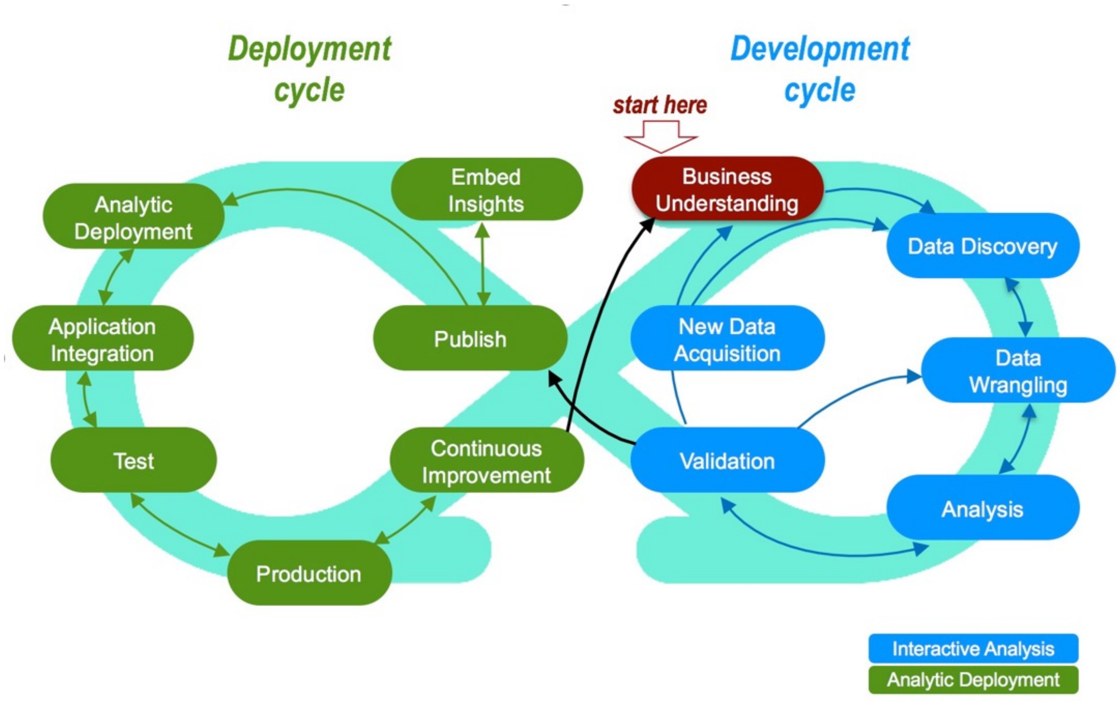
- A patient has a 20% chance of being readmitted to hospital within 30 days
- A customer has a 33% likelihood of not renewing a contract
- 213 transactions are flagged as ‘anomalous’
- A student has a 62% chance of accepting an offer of a university place
- 1,500 guests are estimated to be unsatisfied with their previous stay at a hotel

Given that…

- Decisions drive actions, and actions have consequences and therefore costs
- The context in which these new actions are taken is of critical importance. Especially if:
  - They are simply unfeasible
  - They impact others in the organisation
  - They use up existing resources
  - They are not monitored or measured
  - They are not equitable or fair
- This is why effective planning is so important and *why a number of existing methodologies exist to help*

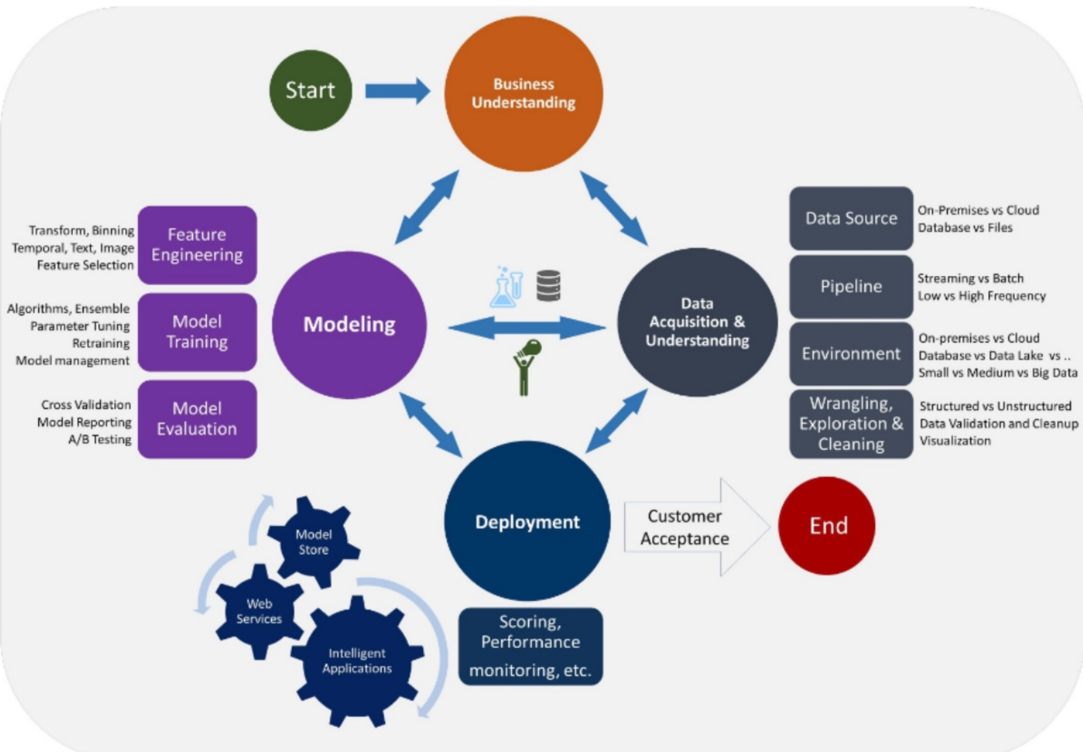
Existing Methodologies

- IBM's Analytics Solution Unified Method (ASUM-DM) methodology for developing Predictive Analytics applications



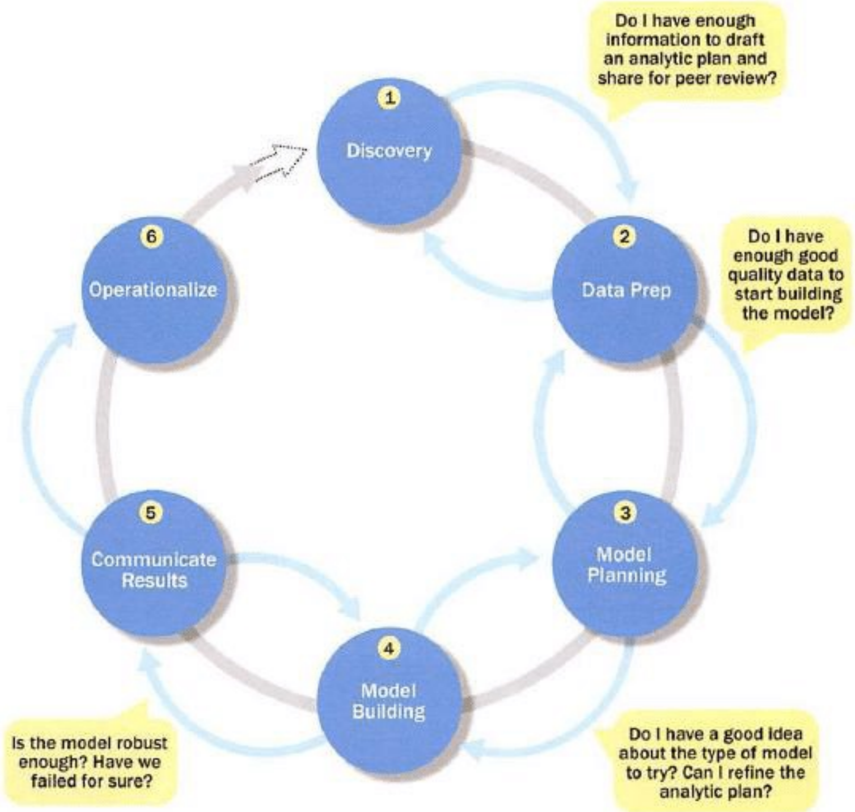
Existing Methodologies

- Microsoft’s Team Data Science Process (TDSP)



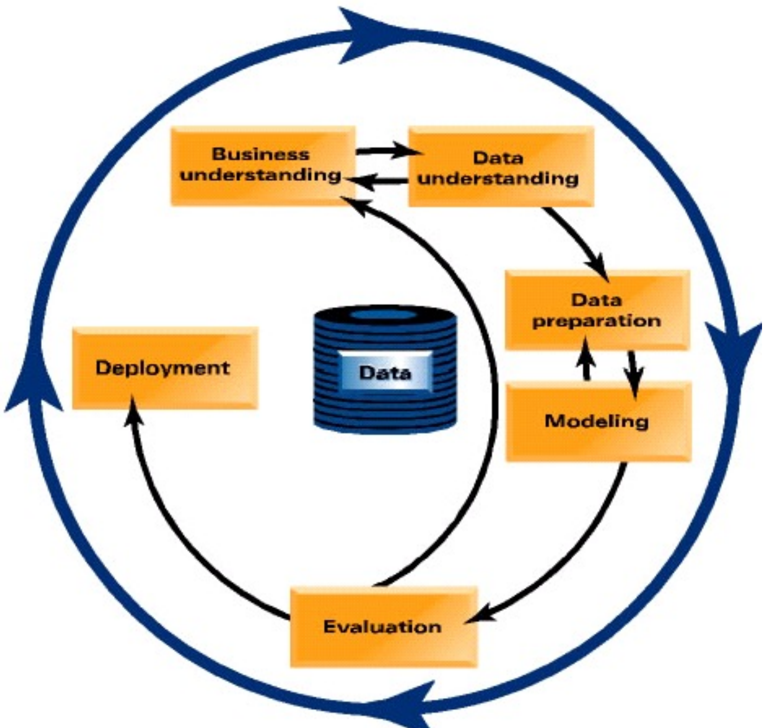
Existing Methodologies

- EMC's Data Analytics Lifecycle



Existing Methodologies

- Non-proprietary Cross-Industry Standard Process for Data Mining (CRISP-DM)



For a start, it's not all about Data Scientists…

Typical Project Roles

Business
- Domain
- Subject Matter

Analytical
- Methodologies
- What to use when

Data
- Management
- Structure

Technology
- Integration
- Building apps

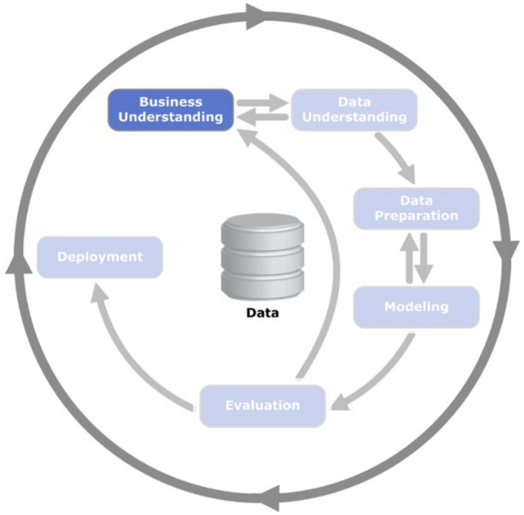
Practical Questions

- Why this particular area of focus?
- How has this been measured thus far?
- What metrics are we trying to hit?
- Who is impacted?
- How will it be tested?
- What are the costs associated with errors?
- What will be done differently as a result of the application?
- How will know you if it worked?
- How will it be updated or maintained?

Business Understanding

Business Understanding

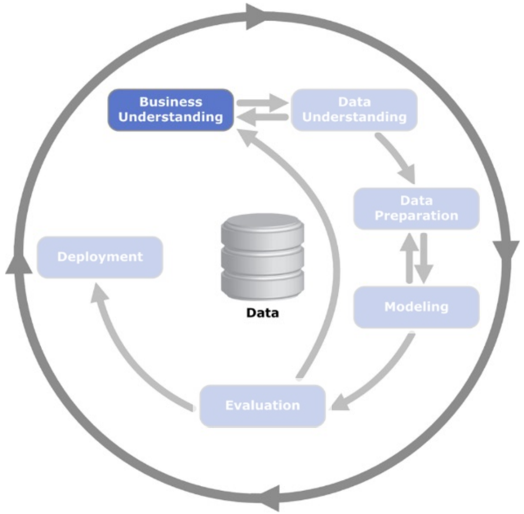
- Often the most overlooked part of the project
- Provides the context and reasoning for the initiative
- Consists of a written and agreed plan with documented and measurable objectives
- Effective Predictive Analytics projects are “front-loaded” with a solid Business Understanding



It is critical that everyone understands what a successful output from the project will look like and what will happen differently as a result of it

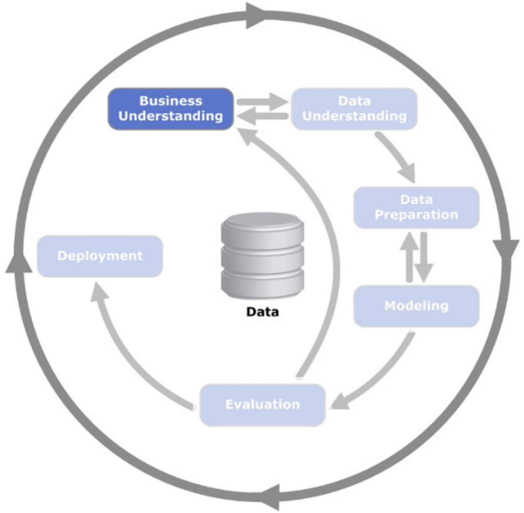
Poorly Documented Business Understanding

- "Aim of the project is to build an accurate model to predict subscribers at risk of cancelling their contracts"
- "By identifying at risk subscribers, effective action in the form of additional offers can be taken to reduce the likelihood of churn"



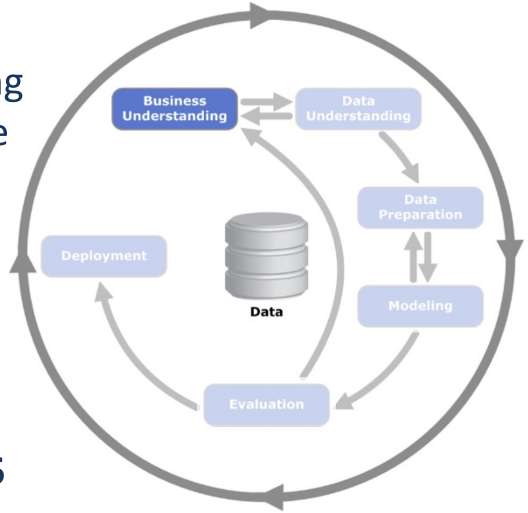
Well Documented Business Understanding

- Primary Business objective: **Increasing customer loyalty**
- "…stipulated as a strategic goal for the company. It's recognised that costs associated with customer acquisition have been rising in recent years and that increasing pressure from competitor activity has led to a slowdown in the growth of market share". **So this is regarded as a *valuable objective*.**
- "Due to the contractual basis of our business model, we can identify precisely when new customers are acquired and when they cease to transact with us". **So this is a *measurable outcome***.
- "Previous analysis using satisfaction surveys has shown that if we are able to identify customers likely to cancel their contracts three months before their termination date, we can persuade around 50% of them to renew their contracts with us". **So this can be regarded as an *actionable result*.**



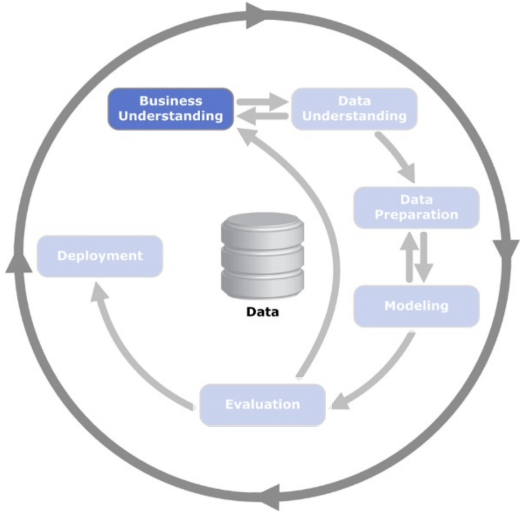
Well Documented Business Understanding

- CRISP-DM dictates that a key output from the Business Understanding task of 'Assessing your situation' should be to figure out the rationale for the application in terms of the financial consequences
- The average cost incurred with persuading customers to renew an existing contract is **$35** (let's assume this is only incurred by those who agree to renew)
- The average annual revenue received from a contract is **$132**
- The cost per head of acquiring and onboarding new customers is **$45**



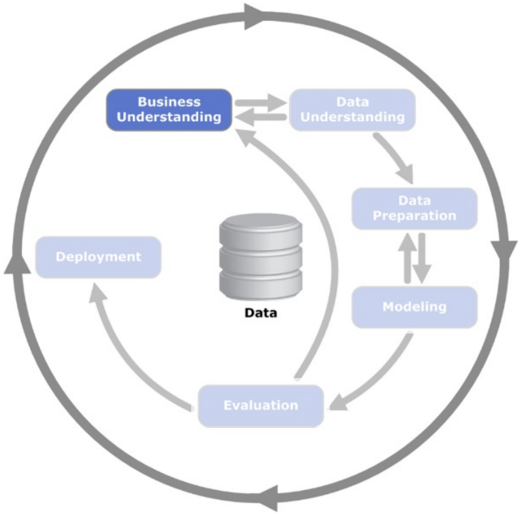
Well Documented Business Understanding

- If we assume that it only takes one month to replace a customer, then the lost revenue on average is only $11
- However, to replace the customer, the acquisition and onboarding costs mean that this value jumps to $56
- If the company is losing 100,000 customers a year (a not unreasonable number) then the total costs are **$5.6 million**
- Let's assume a modest-performing model identifies what it thinks are the top 30,000 customers who are likely to leave (or 'churn') annually. So that is only 30% of the 100,000 churners



Well Documented Business Understanding

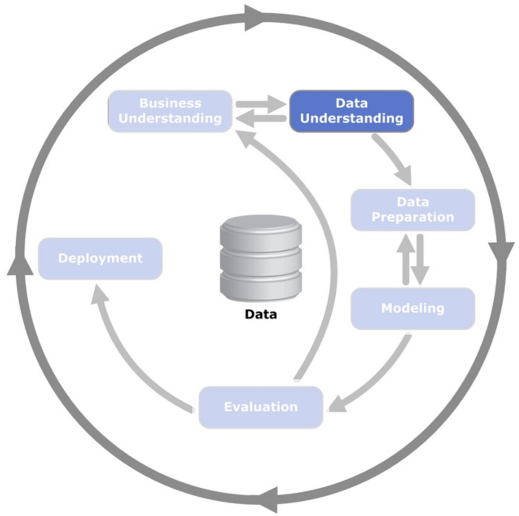
- Of course, we can't assume that the model is completely accurate, so let's assume that it's only right 2 out of 3 times. Meaning that we may only identify 20,000 customers who will churn annually.
- But we can only hope to persuade 50% of them to remain customers. So that's 10,000 customers we have retained and 90,000 customers who cancel their contracts
- It now costs the company $5.04 million to replace the lost customers and they will incur additional retention costs of $350K to persuade 10,000 customers to renew their contracts bringing the total costs to $5.39 million.
- This represents a relatively modest but worthwhile *cost reduction of $210K*.
- Crucially though, in doing so they will have managed to retain 10,000 customers with total annual revenues of $1.35 million.

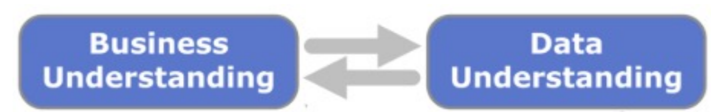


Data Understanding

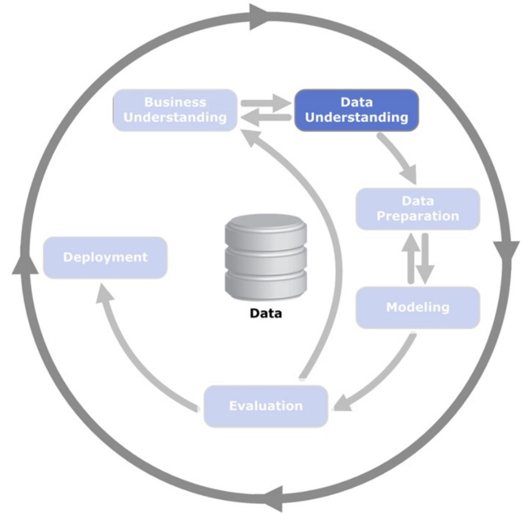
Data Understanding

- This phase asks the question "From a data resource standpoint, what would it take to make this happen?"
- How far back can we go? E.g. has the business model changed recently so that data from over a year ago is no longer relevant?
- There are a number of potential sources of data:
  - Descriptive/Demographic
  - Transactional/Financial
  - Usage/Consumption
  - Problems/Errors
  - Interaction/Contacts
  - Evaluation/Feedback





Data Understanding



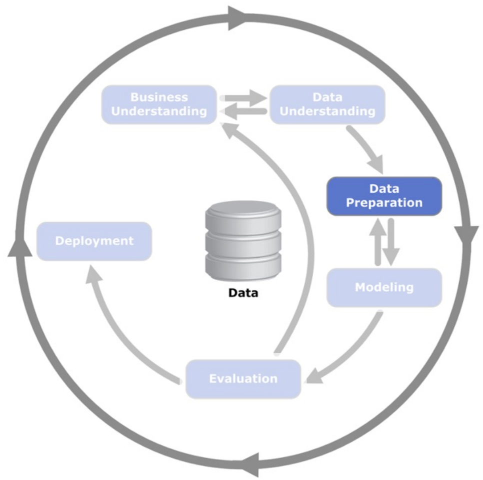
- Do any of the files contain multiple rows per customer (e.g. individual transactions)? Are there any fields that are irrelevant or unclear?
- Are there any highly skewed distributions that can affect summary measures like averages? Do the files match correctly?
- Are there duplicate customer IDs? How complete are the data in terms of missing values?
- Are there fields that contain a large number of categories (e.g. many product codes)

Field	Sample Graph	Measurement	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
Customer_ID		Nominal	1	889366	--	--	--	889366	889366
Age		Continuo...	5	111	42.626	14.481	-0.013	--	884133
Annual_Salary		Continuo...	-9.000	2600594.700	38591.354	19137.216	3.644	--	876263
Gender		Nominal	--	--	--	--	--	5	885009
Product_Code		Nominal	--	--	--	--	--	99	889366

SMARTVISION
Europe

Data Preparation

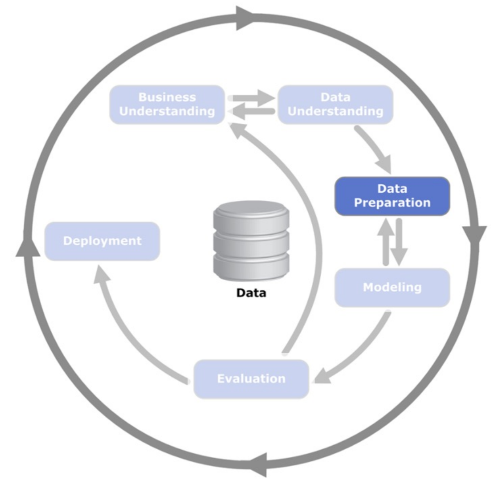
Data Preparation



- Nothing *devours* project time like Data Preparation
- It's estimated that 50% to 70% of allocated days can be used up simply preparing the data for modelling
- The previous two phases allow the project team to imagine what a *single row of data* may need to contain in order to generate the new outcome/prediction
- The data usually weren't collected with modelling in mind. So these resources may require *a lot* preparation to make the application work well

Customer Details	Customer History	Payment History
Gender	Tenure	Number of Missed Payments
Age group	Total Service Usage	Current Payment Method
Marital Status	Average Service usage last month	Card Expiry Date
Postcode	Average Service usage last 3 months	Bonus Content Last 3 Months
License Type	Overall Average service usage	Bonus Content Last 6 Months
Device (Android, IOS)	Complaints	Contract Renewal

Data Preparation



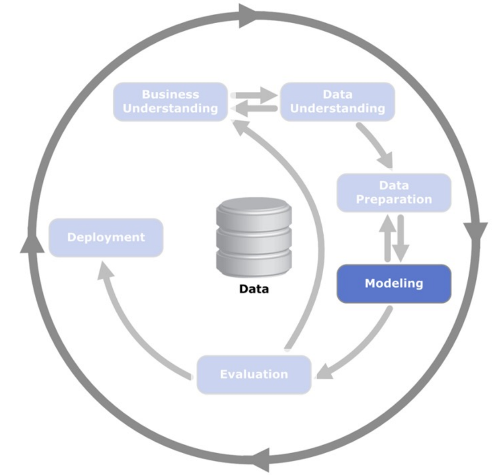
- Typical Data Preparation tasks include:
  - Cleaning and imputing values
  - Merging and appending separate files
  - Aggregation (creating summarised files)
  - Filtering out irrelevant rows
  - Transforming and renaming fields
  - Transposing data (e.g. rows become fields)
  - Creating new fields – recategorizing, calculating elapsed time, deriving ratios, creating banded values, standardising scales

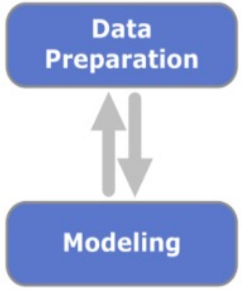
Experienced data analysts know that developing 'good models' is more easily done using well-prepared data than by relying on special algorithms or tuning model parameters.

Modelling

Modelling

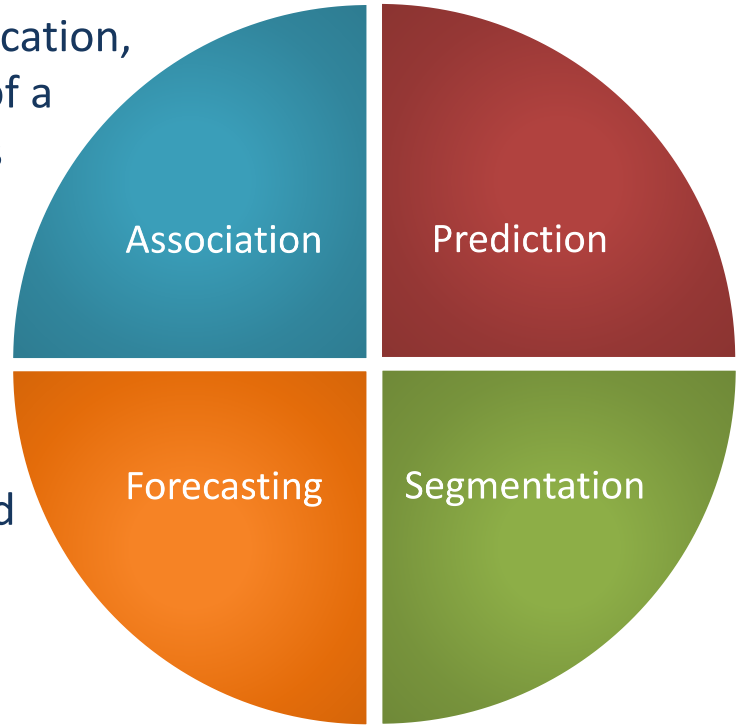
- Sometimes this is the most straightforward part of the entire project
- If the project team have been diligent and thorough in the previous phases, by this stage the data either ‘talks’ or it doesn’t
- Note that CRISP-DM indicates there may be a requirement to iterate back and forth between the Data Preparation and Modelling phases

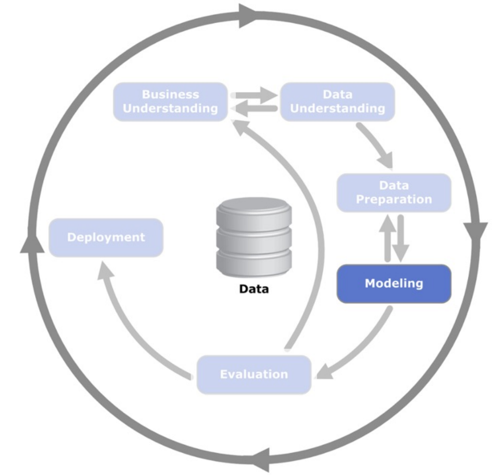




Modelling

- Depending on the application, the model can be one of a range of standard types

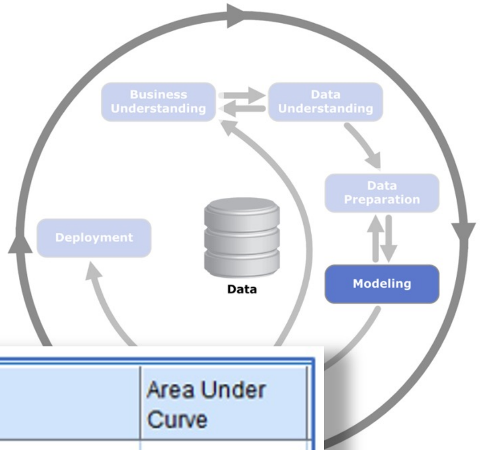




- These techniques may be statistical, rule-based or based on pure machine learning algorithms

Modelling

- It’s possible that a range of models could be developed for comparison
- Different techniques
- Different model parameters
- Different subsets of the data

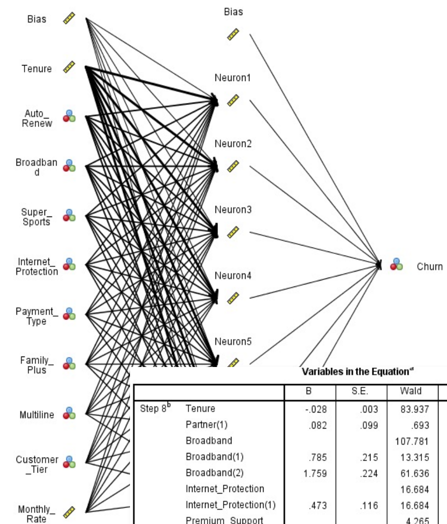


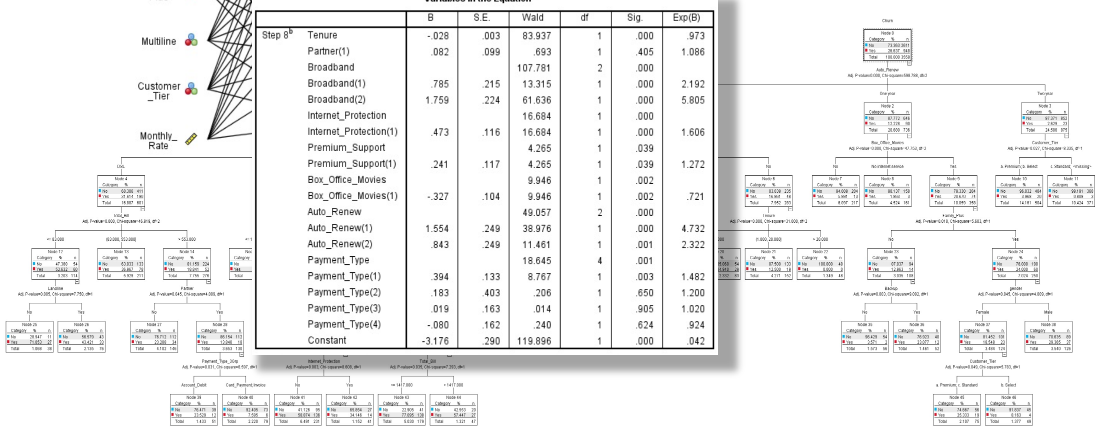
Graph	Model	Area Under Curve
	Logistic regres...	0.846
	Neural Net 1	0.837
	CHAID 1	0.835
	Bayesian Netw...	0.832
	C5 1	0.812

Evaluation

Evaluation

- What makes for a “good” model?
  - Accuracy
  - Interpretability
  - Stability
  - Coherence
  - Simplicity
  - Performance



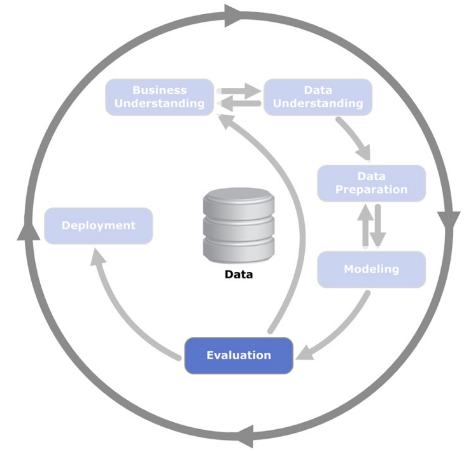


Variables in the Equation[a]

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 8[b]	Tenure	-.028	.003	83.937	1	.000	.973
	Partner(1)	.082	.099	.693	1	.405	1.086
	Broadband			107.781	2	.000	
	Broadband(1)	.785	.215	13.315	1	.000	2.192
	Broadband(2)	1.759	.224	61.636	1	.000	5.805
	Internet_Protection			16.684	1	.000	
	Internet_Protection(1)	.473	.116	16.684	1	.000	1.606
	Premium_Support			4.265	1	.039	
	Premium_Support(1)	.241	.117	4.265	1	.039	1.272
	Box_Office_Movies			9.946	1	.002	
	Box_Office_Movies(1)	-.327	.104	9.946	1	.002	.721
	Auto_Renew			49.057	2	.000	
	Auto_Renew(1)	1.554	.249	38.976	1	.000	4.732
	Auto_Renew(2)	.843	.249	11.461	1	.001	2.322
	Payment_Type			18.645	4	.001	
	Payment_Type(1)	.394	.133	8.767	1	.003	1.482
	Payment_Type(2)	.183	.403	.206	1	.650	1.200
	Payment_Type(3)	.019	.163	.014	1	.905	1.020
	Payment_Type(4)	-.080	.162	.240	1	.624	.924
	Constant	-3.176	.290	119.896	1	.000	.042

Evaluation

- Which model is best?



Model A: Actual Churn by Predicted Churn

			Churn_Predicted No	Churn_Predicted Yes
Churn_Actual	**No**	Frequency	2330	278
		Percent Correct	**89.3%**	10.7%
	Yes	Frequency	473	493
		Percent Correct	49.0%	**51.0%**

Model B: Actual Churn by Predicted Churn

			Churn_Predicted No	Churn_Predicted Yes
Churn_Actual	**No**	Frequency	2129	479
		Percent Correct	**81.6%**	18.4%
	Yes	Frequency	307	659
		Percent Correct	31.8%	**68.2%**

Model C: Actual Churn by Predicted Churn

			Churn_Predicted No	Churn_Predicted Yes
Churn_Actual	**No**	Frequency	1919	689
		Percent Correct	**73.6%**	26.4%
	Yes	Frequency	214	752
		Percent Correct	22.2%	**77.8%**

Evaluation

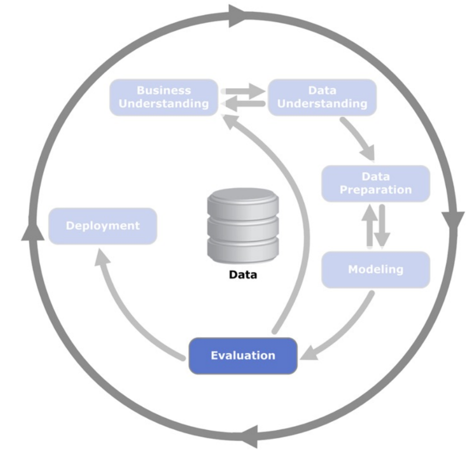
- What are the *costs* of misclassifying cases?

Model A: Actual Churn by Predicted Churn

			Churn_Predicted: No	Churn_Predicted: Yes
Churn_Actual	**No**	Frequency	2330	False 78
		Percent Correct	89.3%	Positive 10.7%
	Yes	Frequency	False 73	493
		Percent Correct	Negative 49.0%	51.0%

Evaluation

- What are the *costs* of misclassifying cases?



Model A: Actual Churn by Predicted Churn

			Churn_Predicted	
			No	**Yes**
Churn_Actual	**No**	Frequency	2330	278
		Percent Correct	**89.3%**	10.7%
	Yes	Frequency	473	493
		Percent Correct	49.0%	**51.0%**

Model B: Actual Churn by Predicted Churn

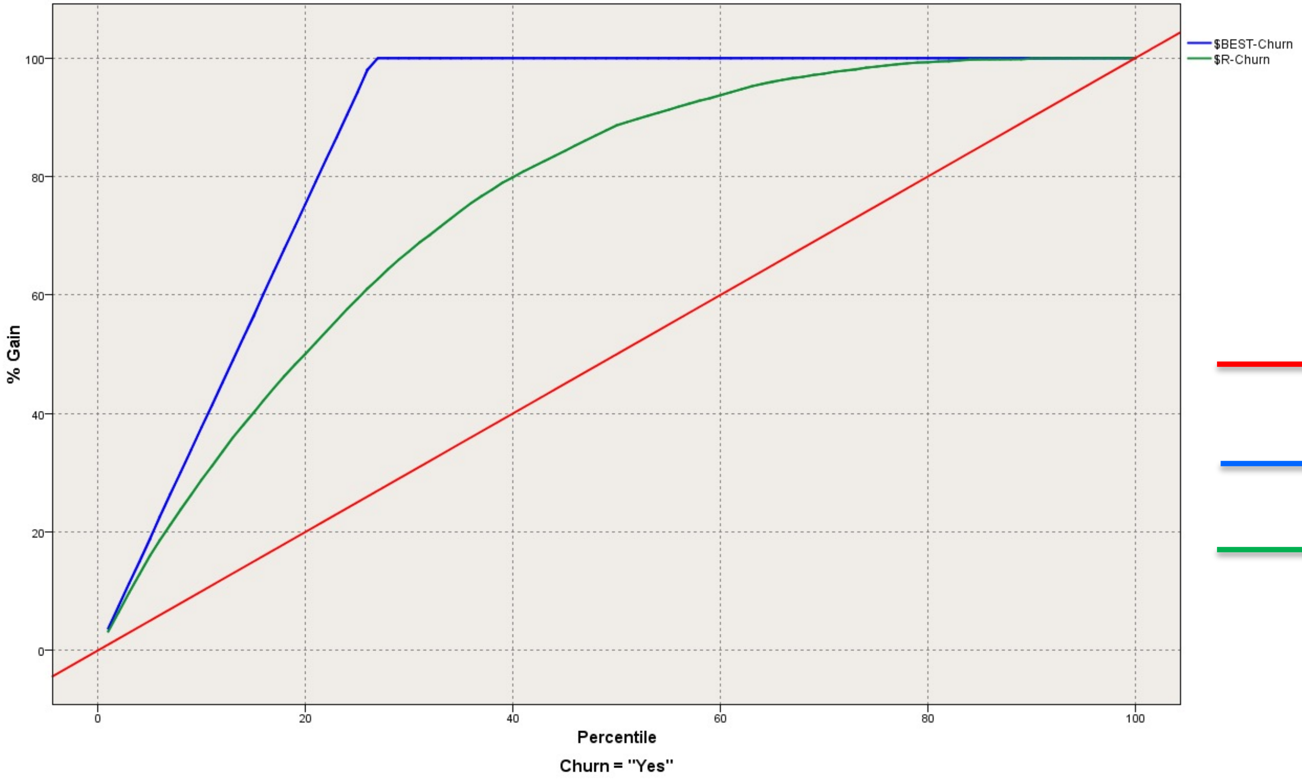
			Churn_Predicted	
			No	**Yes**
Churn_Actual	**No**	Frequency	2129	479
		Percent Correct	**81.6%**	18.4%
	Yes	Frequency	307	659
		Percent Correct	31.8%	**68.2%**

Model C: Actual Churn by Predicted Churn

			Churn_Predicted	
			No	**Yes**
Churn_Actual	**No**	Frequency	1919	689
		Percent Correct	**73.6%**	26.4%
	Yes	Frequency	214	752
		Percent Correct	22.2%	**77.8%**

Evaluation

- Classification models act like data filters



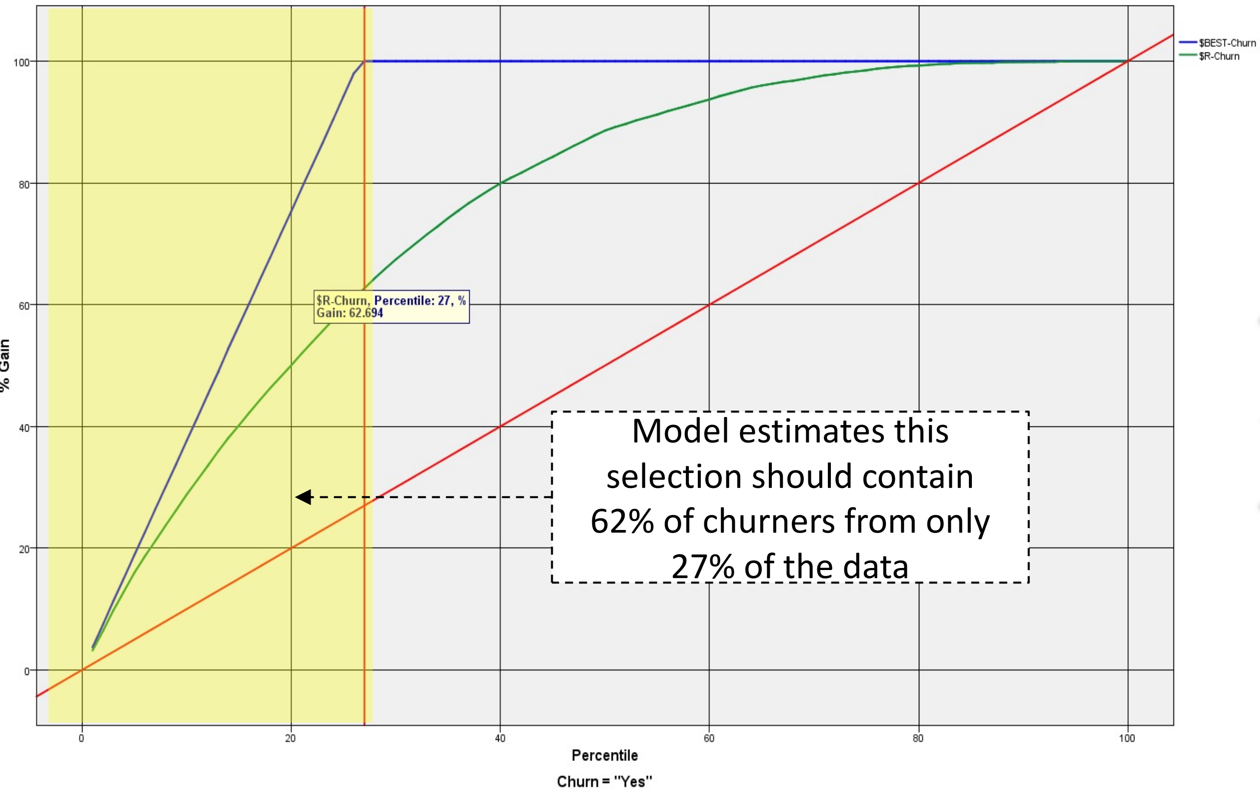
Random performance

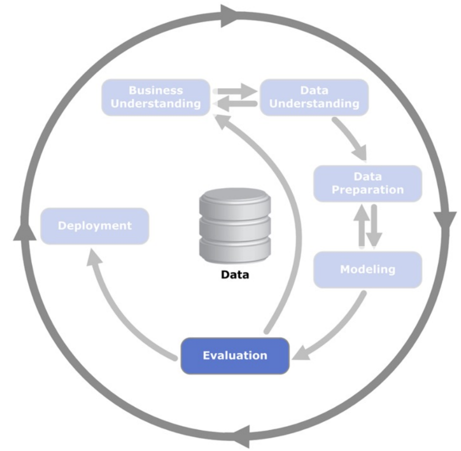
Perfect performance

Model performance

Evaluation

- Classification models act like data filters





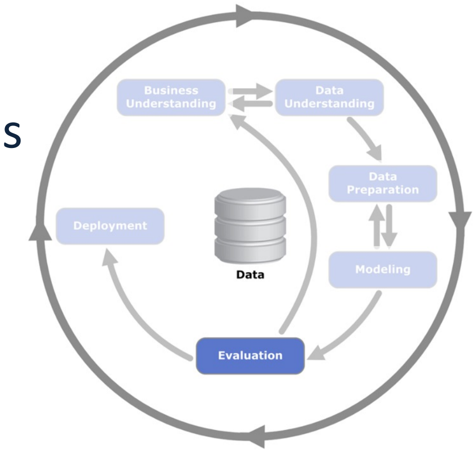
Random performance

Perfect performance

Model performance

Evaluation

- Compare model performance on ‘Train’ and ‘Test’ samples



Graph	Model	Area Under Curve
	CHAID 1	0.855
	Logistic regres...	0.849
	Neural Net 1	0.848
	Bayesian Netw...	0.838
	C5 1	0.826

Training Sample Performance

Graph	Model	Area Under Curve
	Logistic regres...	0.846
	Neural Net 1	0.837
	CHAID 1	0.835
	Bayesian Netw...	0.832
	C5 1	0.812

Testing Sample Performance

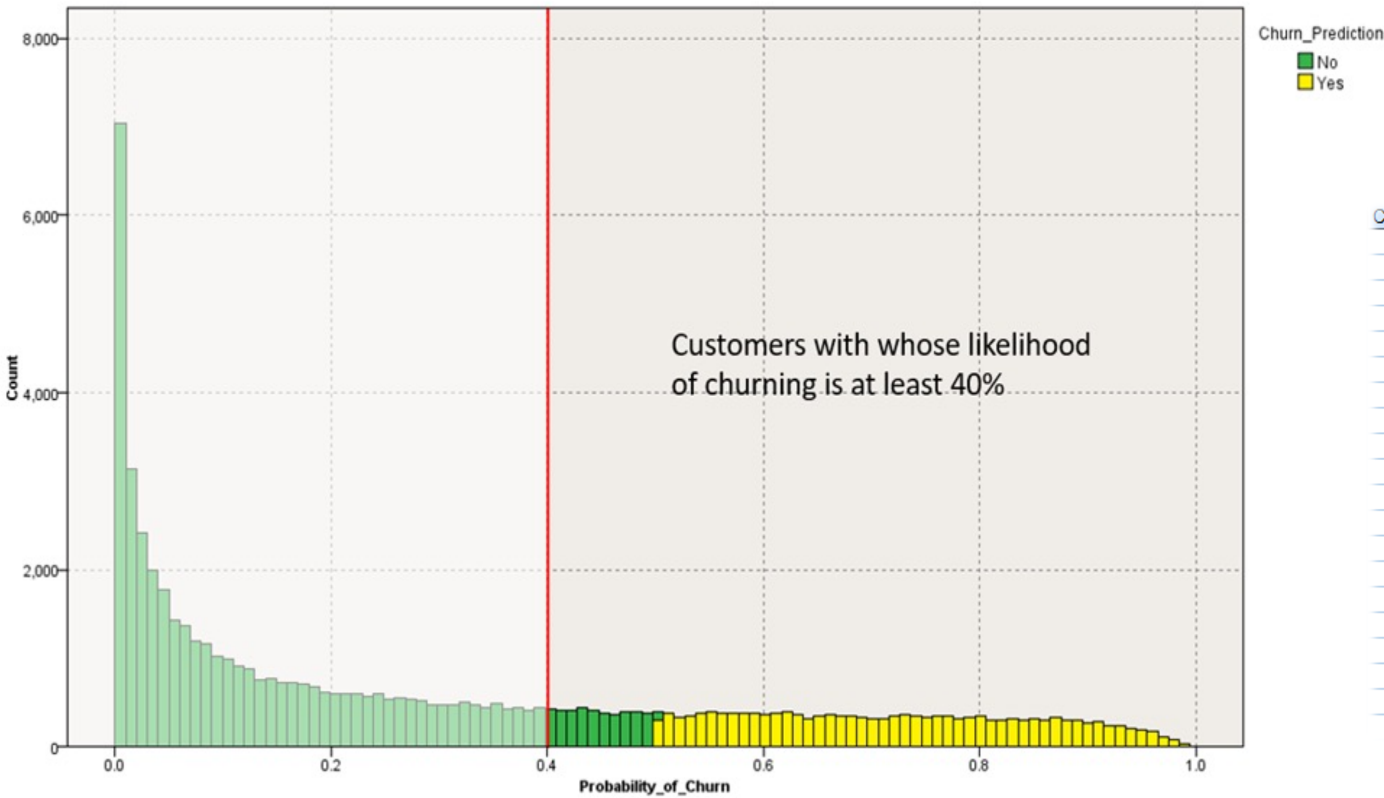
Evaluation

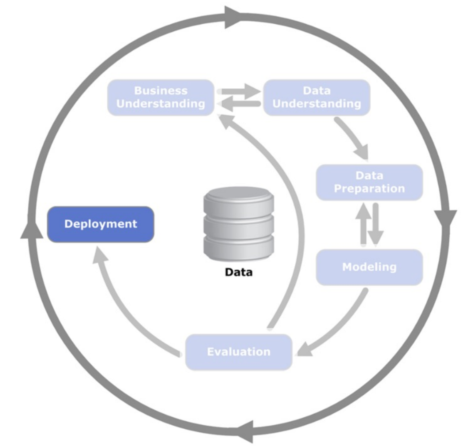
- No single way to evaluate model performance
- This is why defining the model performance criteria in terms of the Business Understanding is so important
- A simple approach might be to say “we aim to select 30,000 customers with a sufficiently high risk of churning that we would expect at least 20,000 of them to actually churn”
- You can then select the model that is most likely to meet this criterion as well as any other criteria such as transparency, simplicity and coherence.

Deployment

Deployment

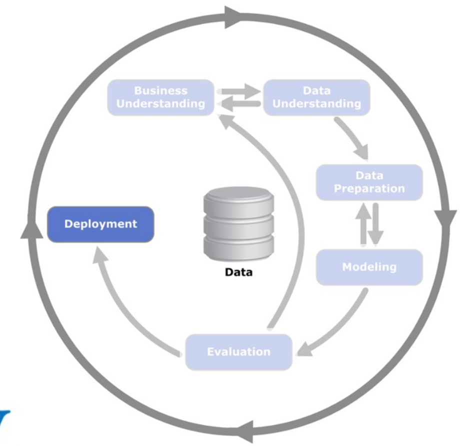
- Deployment might take the form of simply choosing people based on a model score



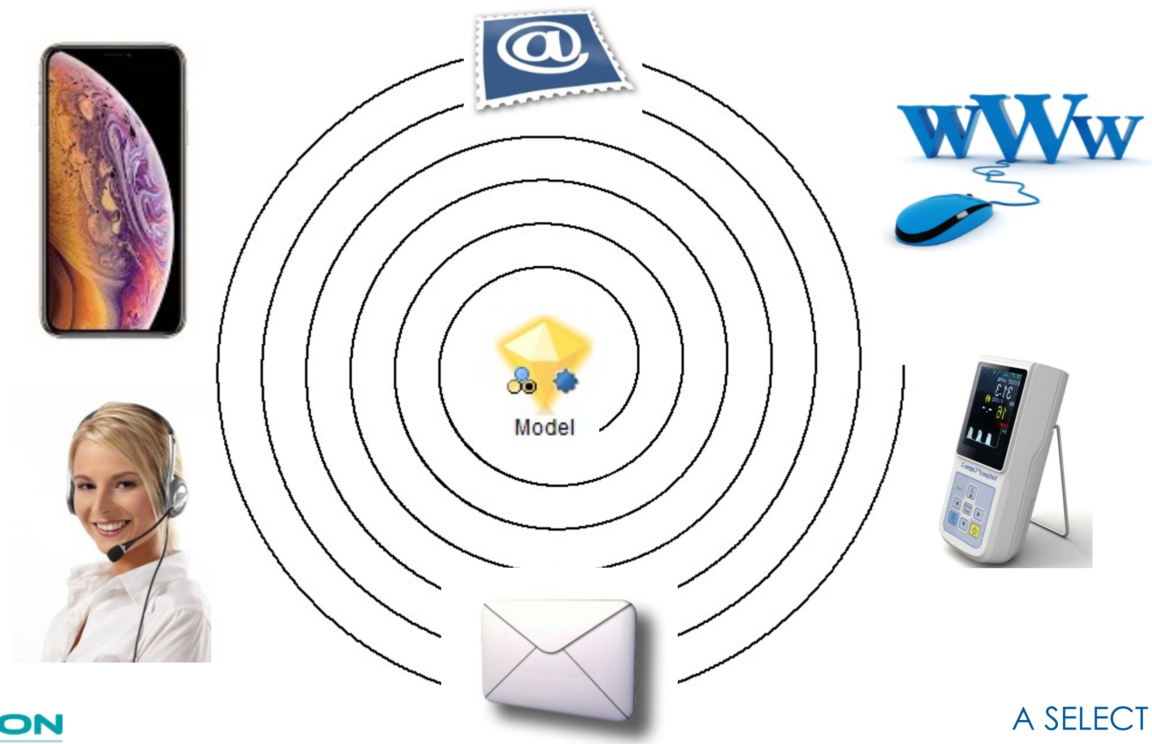


CustomerID	Churn_Prediction	Probability_of_Churn
335.000	No	0.402
6814.000	Yes	0.513
3801.000	Yes	0.607
5653.000	Yes	0.538
784.000	Yes	0.608
6778.000	Yes	0.860
5862.000	Yes	0.928
2971.000	Yes	0.928
5268.000	Yes	0.695
6856.000	No	0.465
5912.000	Yes	0.580
1524.000	Yes	0.779
801.000	Yes	0.888
5756.000	Yes	0.692
1494.000	Yes	0.793
6026.000	Yes	0.846
2031.000	Yes	0.909
692.000	No	0.476
2195.000	Yes	0.620
1401.000	Yes	0.840

Deployment happens in the real world

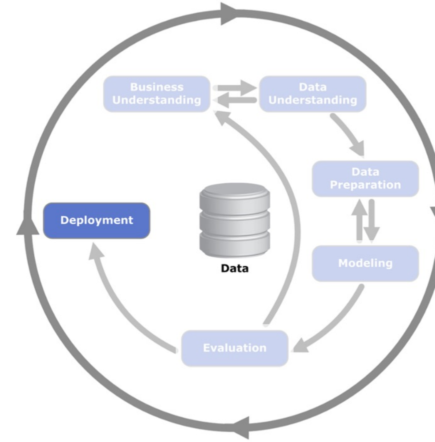


- How the models scores are used depends on the context on the application



Deployment

- *Proving* that the application works may mean that we have to think of the entire exercise as an experiment with testable results



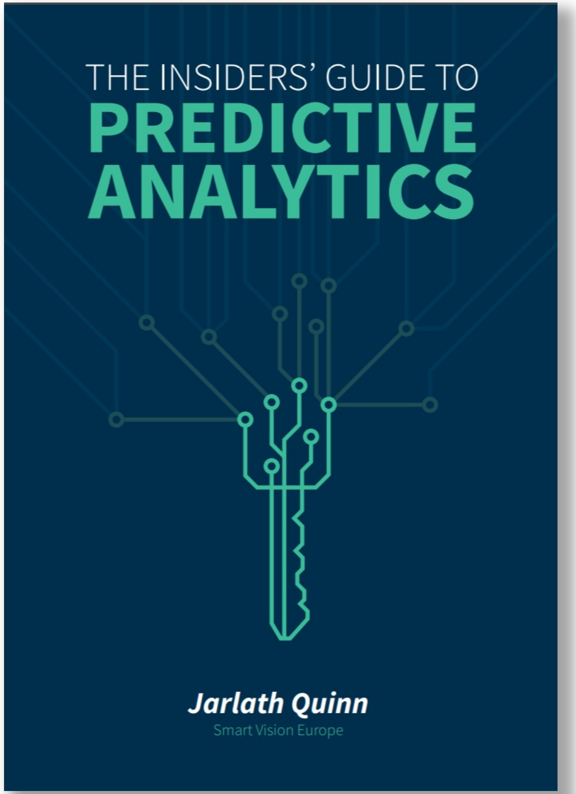
What do we (Smart Vision) talk about when we're discussing a prospective Predictive Analytics application?

1. Why do they want to do this?
2. What will it take to make it work?
3. What does 'good' look like?
4. How will we know it worked?

What do we not talk about when we're discussing a prospective Predictive Analytics application?

1. Algorithms

Download our new e-book for free



The insider's guide to predictive analytics

£0.00

- 1 + **Add to basket**

Category: books

SMARTVISION Europe

A SELECT INTERNATIONAL COMPANY

Working with Smart Vision Europe Ltd.

- **Sourcing Software**
  - You can buy your analytical software from us often with discounts
  - Assist with selection, pilot, implementation & support of analytical tools
  - http://www.sv-europe.com/buy-spss-online/
- **Training and Consulting Services**
  - Guided consulting & training to develop in house skills
  - Delivery of classroom training courses / side by side training support
  - Identification & recruitment of analytical skills into your organisation
- **Advice and Support**
  - offer 'no strings attached' technical and business advice relating to analytical activities
  - Technical support services

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