Fundamentals of Machine Learning for Predictive Data Analytics

Chapter 2: Data to Insights to Decisions

John Kelleher and Brian Mac Namee and Aoife D'Arcy

john.d.kelleher@dit.ie brian.macnamee@ucd.ie aoife@theanalyticsstore.com

Converting Business Problems into Analytics Solutions

- Converting a business problem into an analytics solution involves answering the following key questions:
  1. What is the business problem?
  2. What are the goals that the business wants to achieve?
  3. How does the business currently work?
  4. In what ways could a predictive analytics model help to address the business problem?

Case Study: Motor Insurance Fraud

In spite of having a fraud investigation team that investigates up to 30% of all claims made, a motor insurance company is still losing too much money due to fraudulent claims.

- What predictive analytics solutions could be proposed to help address this business problem?

- Potential analytics solutions include:
  - Claim prediction
  - Member prediction
  - Application prediction
  - Payment prediction

Assessing Feasibility

- Evaluating the feasibility of a proposed analytics solution involves considering the following questions:
    1. Is the data required by the solution available, or could it be made available?
    2. What is the capacity of the business to utilize the insights that the analytics solution will provide?

- What are the data and capacity requirements for the proposed Claim Prediction analytics solution for the motor insurance fraud scenario?

- What are the data and capacity requirements for the proposed Claim Prediction analytics solution for the motor insurance fraud scenario?

Case Study: Motor Insurance Fraud

[Claim prediction]
Data Requirements: A large collection of historical claims marked as *'fraudulent'* and *'non-fraudulent'*. Also, the details of each claim, the related policy, and the related claimant would need to be available.
Capacity Requirements: The main requirement is that a mechanism could be put in place to inform claims investigators that some claims were prioritized above others. This would also require that information about claims become available in a suitably timely manner so that the claims investigation process would not be delayed by the model.

Designing the Analytics Base Table

- The basic structure in which we capture historical datasets is the **analytics base table** (**ABT**)

Descriptive Features						Target Feature
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----
----	----	----	----	----	----	----

Figure: The general structure of an **analytics base table**—descriptive features and a target feature.

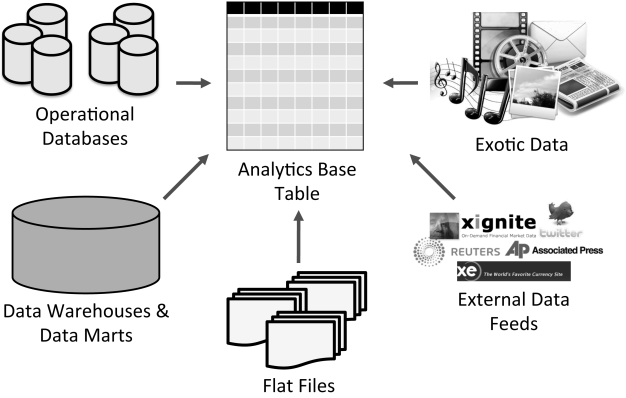


Figure: The different data sources typically combined to create an analytics base table.

- The **prediction subject** defines the basic level at which predictions are made, and each row in the ABT will represent one instance of the prediction subject—the phrase **one-row-per-subject** is often used to describe this structure.
- Each row in an ABT is composed of a set of descriptive features and a target feature.
- Defining features can be difficult!

- A good way to define features is to identify the key **domain concepts** and then to base the features on these concepts.

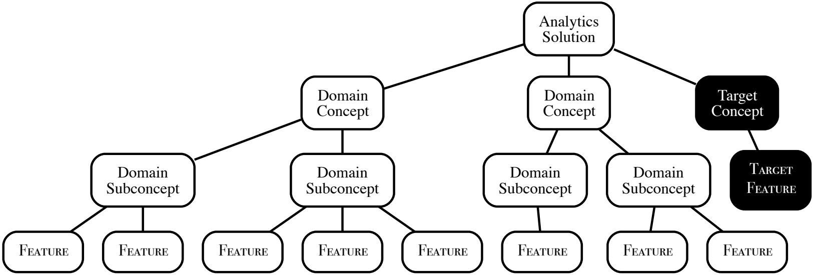


Figure: The hierarchical relationship between an analytics solution, domain concepts, and descriptive features.

- There are a number of general domain concepts that are often useful:
  - Prediction Subject Details
  - Demographics
  - Usage
  - Changes in Usage
  - Special Usage
  - Lifecycle Phase
  - Network Links

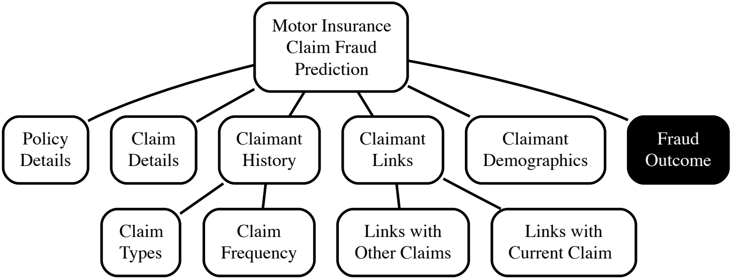


Figure: Example domain concepts for a motor insurance fraud claim prediction analytics solution.

Designing & Implementing Features

- Three key data considerations are particularly important when we are designing features.
  - **Data availability**
  - **Timing**
  - **Longevity**

Ordinal Ordinal Categorical

ID	Name	Date of Birth	Gender	Credit Rating	Country	Salary
0034	Brian	22/05/78	male	aa	ireland	67,000
0175	Mary	04/06/45	female	c	france	65,000
0456	Sinead	29/02/82	female	b	ireland	112,000
0687	Paul	11/11/67	male	a	usa	34,000
0982	Donald	01/12/75	male	b	australia	88,000
1103	Agnes	17/09/76	female	aa	sweden	154,000

Textual Interval Binary Numeric

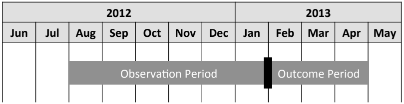
Figure: Sample descriptive feature data illustrating numeric, binary, ordinal, interval, categorical, and textual types.

- The features in an ABT can be of two types:
  - **raw features**
  - **derived features**
- There are a number of common derived feature types:
  - **Aggregates**
  - **Flags**
  - **Ratios**
  - **Mappings**

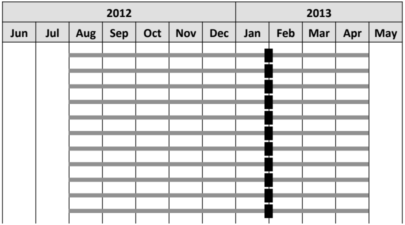
Handling Time

- Many of the predictive models that we build are **propensity models**, which inherently have a temporal element
- For **propensity modeling**, there are two key periods:
  - the **observation period**
  - the **outcome period**

- In some cases the observation and outcome period are measured over the same time for all predictive subjects.



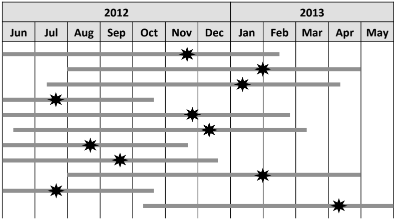
(a) Observation period and outcome period



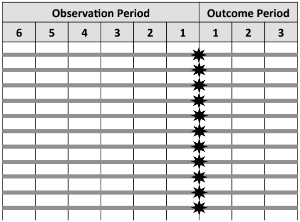
(b) Observation and outcome periods for multiple customers (each line represents a customer)

Figure: Modeling points in time.

- Often the observation period and outcome period will be measured over different dates for each prediction subject.



(a) Actual

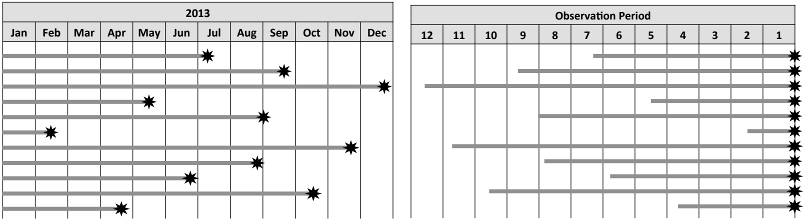


(b) Aligned

Figure: Observation and outcome periods defined by an event rather than by a fixed point in time (each line represents a prediction subject and stars signify events).

Handling Time

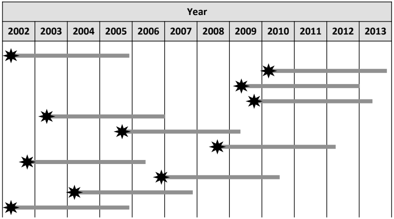
- In some cases only the descriptive features have a time component to them, and the target feature is time independent.



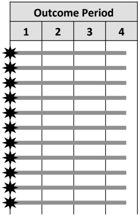
(a) Actual (b) Aligned

Figure: Modeling points in time for a scenario with no real outcome period (each line represents a customer, and stars signify events).

- Conversely, the target feature may have a time component and the descriptive features may not.



(a) Actual



(b) Aligned

Figure: Modeling points in time for a scenario with no real observation period (each line represents a customer, and stars signify events).

Legal Issues

- Data analytics practitioners can often be frustrated by legislation that stops them from including features that appear to be particularly well suited to an analytics solution in an ABT.
- There are significant differences in legislation in different jurisdictions, but a couple of key relevant principles almost always apply.
  1. **Anti-discrimination legislation**
  2. **Data protection legislation**

- Although, data protection legislation changes significantly across different jurisdictions, there are some common tenets on which there is broad agreement which affect the design of ABTs
  - The **collection limitation principle**
  - The **purpose specification principle**
  - The **use limitation principle**

Implementing Features

- Implementing a **derived feature**, however, requires data from multiple sources to be combined into a set of single feature values.
- A few key **data manipulation** operations are frequently used to calculate derived feature values:
  - joining data sources
  - filtering rows in a data source
  - filtering fields in a data source
  - deriving new features by combining or transforming existing features
  - aggregating data sources

Case Study: Motor Insurance Fraud

- What are the observation period and outcome period for the motor insurance claim prediction scenario?

Case Study: Motor Insurance Fraud

- What are the observation period and outcome period for the motor insurance claim prediction scenario?
- The observation period and outcome period are measured over different dates for each insurance claim, defined relative to the specific date of that claim.

Case Study: Motor Insurance Fraud

- What are the observation period and outcome period for the motor insurance claim prediction scenario?
- The observation period and outcome period are measured over different dates for each insurance claim, defined relative to the specific date of that claim.
- The observation period is the time prior to the claim event, over which the descriptive features capturing the claimant's behavior are calculated

Case Study: Motor Insurance Fraud

- What are the observation period and outcome period for the motor insurance claim prediction scenario?
- The observation period and outcome period are measured over different dates for each insurance claim, defined relative to the specific date of that claim.
- The observation period is the time prior to the claim event, over which the descriptive features capturing the claimant's behavior are calculated
- The outcome period is the time immediately after the claim event, during which it will emerge whether the claim is fraudulent or genuine.

Case Study: Motor Insurance Fraud

What features could you use to capture the Claim Frequency domain concept?

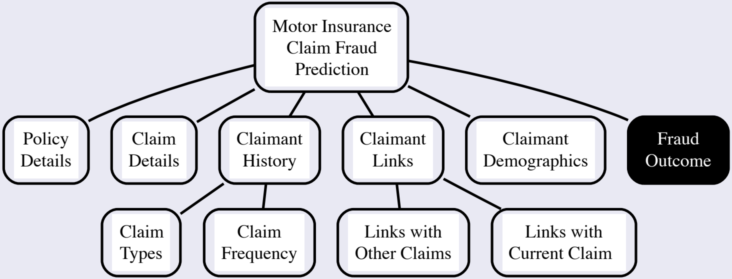


Figure: Example domain concepts for a motor insurance fraud prediction analytics solution.

Case Study: Motor Insurance Fraud

What features could you use to capture the Claim Frequency domain concept?

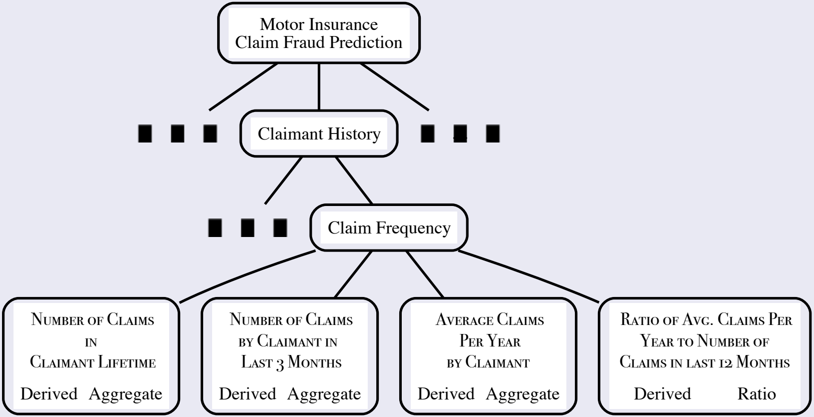


Figure: A subset of the domain concepts and related features for a motor insurance fraud prediction analytics solution.

Case Study: Motor Insurance Fraud

What features could you use to capture the Claim Types domain concept?

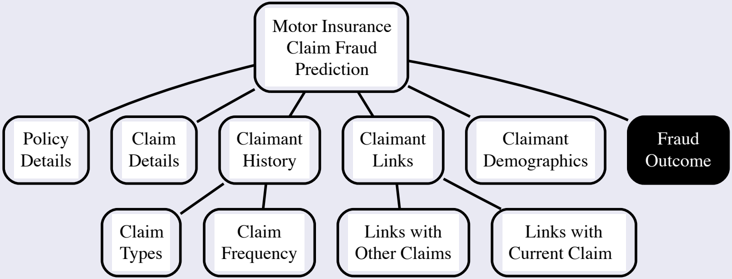


Figure: Example domain concepts for a motor insurance fraud prediction analytics solution.

Case Study: Motor Insurance Fraud

What features could you use to capture the Claim Types domain concept?

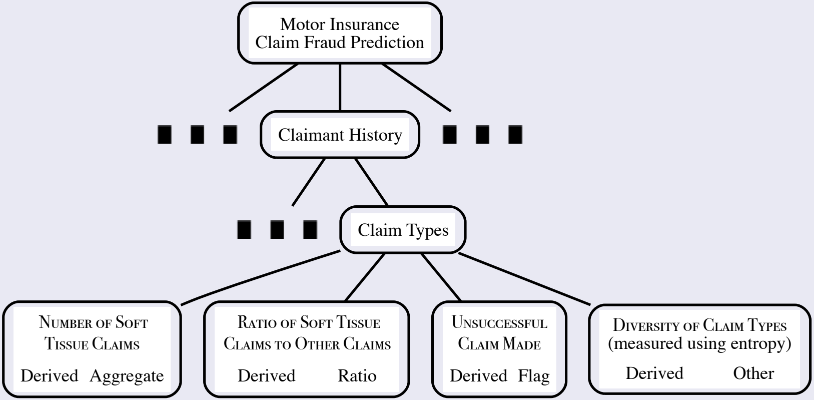


Figure: A subset of the domain concepts and related features for a motor insurance fraud prediction analytics solution.

Case Study: Motor Insurance Fraud

What features could you use to capture the Claim Details domain concept?

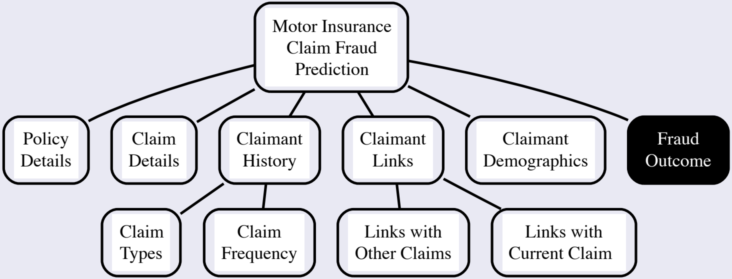


Figure: Example domain concepts for a motor insurance fraud prediction analytics solution.

Case Study: Motor Insurance Fraud

What features could you use to capture the Claim Details domain concept?

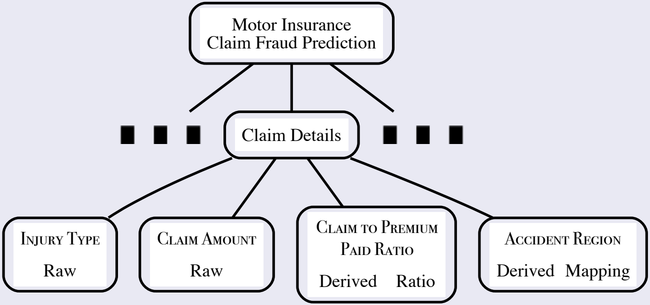


Figure: A subset of the domain concepts and related features for a motor insurance fraud prediction analytics solution.

Case Study: Motor Insurance Fraud

- The following table illustrates the structure of the final ABT that was designed for the motor insurance claims fraud detection solution.
- The table contains more descriptive features than the ones we have discussed
- The table also shows the first four instances.
- If we examine the table closely, we see a number of strange values (for example, $-9\,999$) and a number of missing values—we will return to these in Chapter 3.

Table: The ABT for the motor insurance claims fraud detection solution.

ID	TYPE	INC.	MARITAL STATUS	NUM. CLMNTS.	INJURY TYPE	HOSPITAL STAY	CLAIM AMT.
1	CI	0		2	Soft Tissue	No	1 625
2	CI	0		2	Back	Yes	15 028
3	CI	54 613	Married	1	Broken Limb	No	-9 999
4	CI	0		3	Serious	Yes	270 200
		⋮				⋮	

ID	TOTAL CLAIMED	NUM. CLAIMS	NUM. CLAIMS 3 MONTHS	AVG. CLAIMS PER YEAR	AVG. CLAIMS RATIO	NUM. SOFT TISSUE	% SOFT TISSUE
1	3 250	2	0	1	1	2	1
2	60 112	1	0	1	1	0	0
3	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0
		⋮			⋮		

ID	UNSUCC. CLAIMS	CLAIM AMT. REC.	CLAIM DIV.	CLAIM TO PREM.	REGION	FRAUD FLAG
1	2	0	0	32.5	MN	1
2	0	15 028	0	57.14	DL	0
3	0	572	0	-89.27	WAT	0
4	0	270 200	0	30.186	DL	0
		⋮		⋮		

Summary

- Predictive data analytics models built using machine learning techniques are tools that we can use to help make better decisions within an organization, not an end in themselves.
- It is important to fully understand the business problem that a model is being constructed to address—this is the goal behind *converting business problems into analytics solutions*

- Predictive data analytics models are reliant on the data that is used to build them—the **analytics base table** (**ABT**).
- The first step in designing an ABT is to decide on the **prediction subject**.
- An effective way in which to design ABTs is to start by defining a set of **domain concepts** in collaboration with the business, and then designing **features** that express these concepts in order to form the actual ABT.

- Features (both descriptive and target) are concrete numeric or symbolic representations of domain concepts.
- It is useful to distinguish between **raw features** that come directly from existing data sources and **derived features** that are constructed by manipulating values from existing data sources.
- Common manipulations used in this process include aggregates, flags, ratios, and mappings, although any manipulation is valid.

- The techniques described here cover the **Business Understanding**, **Data Understanding**, and (partially) **Data Preparation** phases of the **CRISP-DM** process.

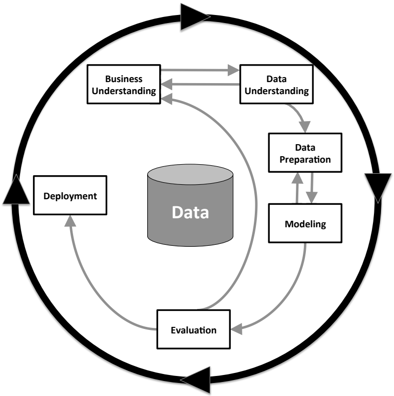


Figure: A diagram of the CRISP-DM process.

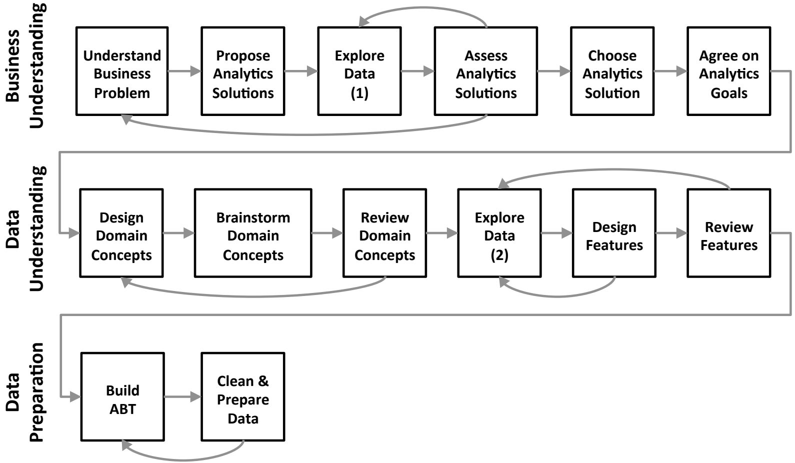


Figure: A summary of the tasks in the Business Understanding, Data Understanding, and Data Preparation phases of the **CRISP-DM** process.