Data to Insights to Decisions

CS576 MACHINE LEARNING

Dr. Jin S. Yoo, Professor
Department of Computer Science
Purdue University Fort Wayne

Reference

- Kelleher et al., Fundamentals of Machine Learning for Predictive Data Analytics, Ch 2

Data to Insight to Decisions

- Part I: Approach to developing **analytics solutions** that address specific **business problems**
  - Analysis of the needs of the business,
  - The data available for use
  - The capacity of the business to use analytics

- Part II: Data structures required to build predictive analytics models – **Analytics based table (ABT)**
  - A set of **domain concepts** that descry be the prediction subject
  - Concrete **descriptive features**

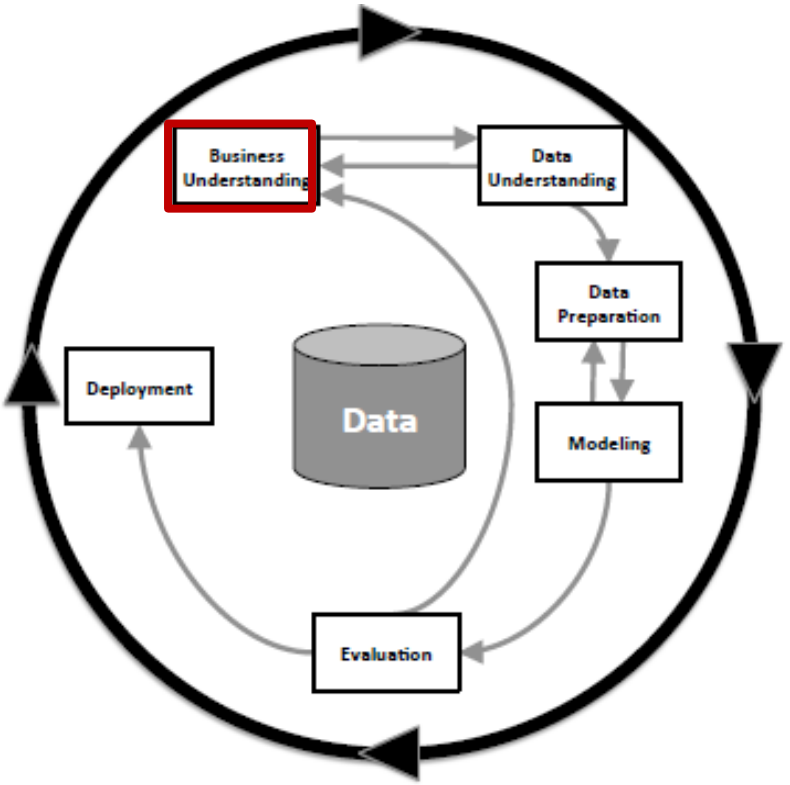
Outline

- Converting Business Problems into Analytics Solutions
- Assessing Feasibility
- Designing the Analytics Base Table
- Designing & Implementing Features
  - Different Types of Data
  - Different Types of Features
  - Handling Time
  - Legal Issues
  - Implementing Features
  - Case Study: Motor Insurance Fraud
- Summary

Converting Business Problems into Analytics Solutions

- A key step in any data analytics project is to **understand the business problem** that the organization wants to solve.
- This defines the **analytics solutions** that the analytics practitioner will set out to build using machine learning.

- Defining the analytics solution is the most important task in the **Business Understanding** phase of the CRISP-DM (Cross Industry Standard Process for Data Mining) process.



Key Questions

1. **What is the business problem? What are the goals that the business wants to achieve?**
   - Unless a project is focused on clearly stated goals, it is unlikely to be successful.
2. **How does the business currently work?**
   - Analytics practitioners must understand enough about a business so that they can converse with partners in the business in a way (i.e., in domain terms) that these business partners understand. **- situational fluency**
3. **In what ways could a predictive analytics model help to address the business problem?**
   - For each proposed solution, describe
     - (1) the predictive model that will be built;
     - (2) how the predictive model will be used by the business;
     - (3) how using the predictive model will help address the original 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

Motor Insurance Fraud – Analytics Examples

- **Claim prediction**
  - Predict the likelihood that an insurance claim is fraudulent
  - *Benefit*: the limited claims investigation time could be targeted at the claims that are most likely to be fraudulent, thereby increasing the number of fraudulent claims detected and reducing the amount of money lost to fraud.
- **Member prediction**
  - Predict the propensity of a member to commit fraud in the near future
  - *Benefit*: By identifying members likely to make fraudulent claims before they make them, the company could save significant amounts of money.

Analytics Examples (Cont.)

- **Application prediction**
  - Predict, at the point of application, the likelihood that a policy someone has applied for will ultimately result in a fraudulent claim
  - *Benefit*: Reduce the number of fraudulent claims and reduce the amount of money they would lose to these claims. Member prediction
- **Payment prediction**
  - Predict the amount most likely to be paid out by an insurance company after having investigated a claim
  - *Benefit*: save on claims investigations and reduce the amount of money paid out on fraudulent claims.

Outline

- Converting Business Problems into Analytics Solutions
- ☞ **Assessing Feasibility**
- Designing the Analytics Base Table
- Designing & Implementing Features
  - Different Types of Data
  - Different Types of Features
  - Handling Time
  - Legal Issues
  - Implementing Features
  - Case Study: Motor Insurance Fraud
- Summary

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?

Case Study: Motor Insurance Fraud

[Claim prediction]

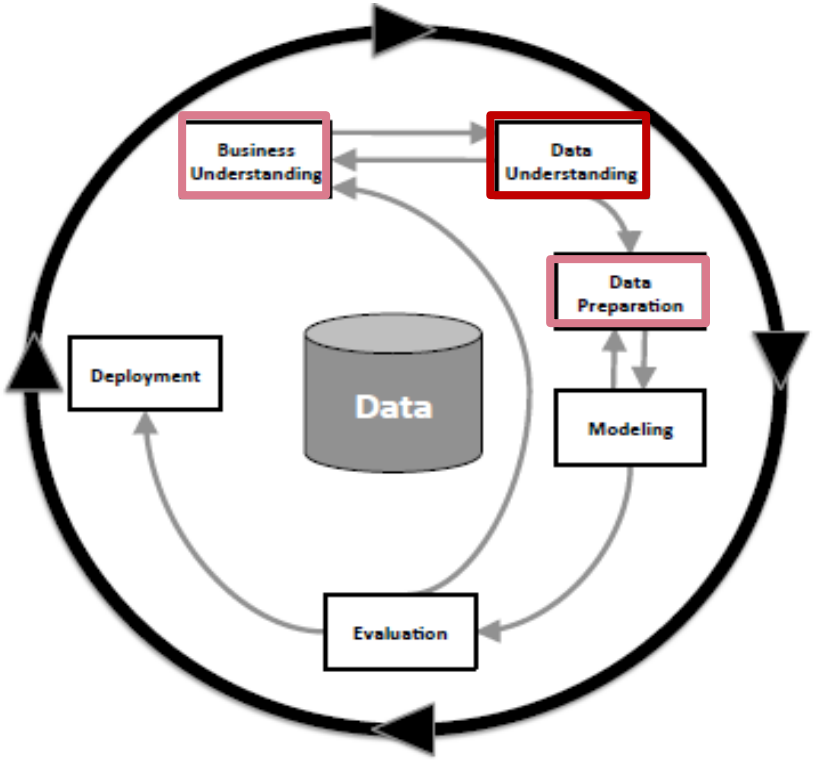
- *Data Requirements*
  - A large collection of historical claims marked as ’fraudulent’ and ’non-fraudulent’.
  - The details of each claim, the related policy, and the related claimant would need to be available.
- *Capacity Requirements*:
  - A mechanism could be put in place to inform claims investigators that some claims were prioritized above others.
  - Information about claims become available in a suitably timely manner so that the claims investigation process would not be delayed by the model.

Outline

- Converting Business Problems into Analytics Solutions
- Assessing Feasibility
- ☞ **Designing the Analytics Base Table**
- Designing & Implementing Features
  - Different Types of Data
  - Different Types of Features
  - Handling Time
  - Legal Issues
  - Implementing Features
  - Case Study: Motor Insurance Fraud
- Summary

Designing the Analytics Base Table

- Develop in response to a business problem, we need to begin to design the data structures that will be used to build, evaluate, and ultimately deploy the model.
- Primarily in **Data Understanding** phase of the CRISP-DM process, and also overlap with **Business Understanding** and **Data Preparation** phases



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

- An analytics base table is a simple, flat, tabular data structure made up of rows and columns.
  - The columns are divided into a set of **descriptive features** and a single **target feature**.
  - Each row contains a value for each descriptive feature and the target feature represents an **instance** about which a prediction can be made.

Raw Data Sources

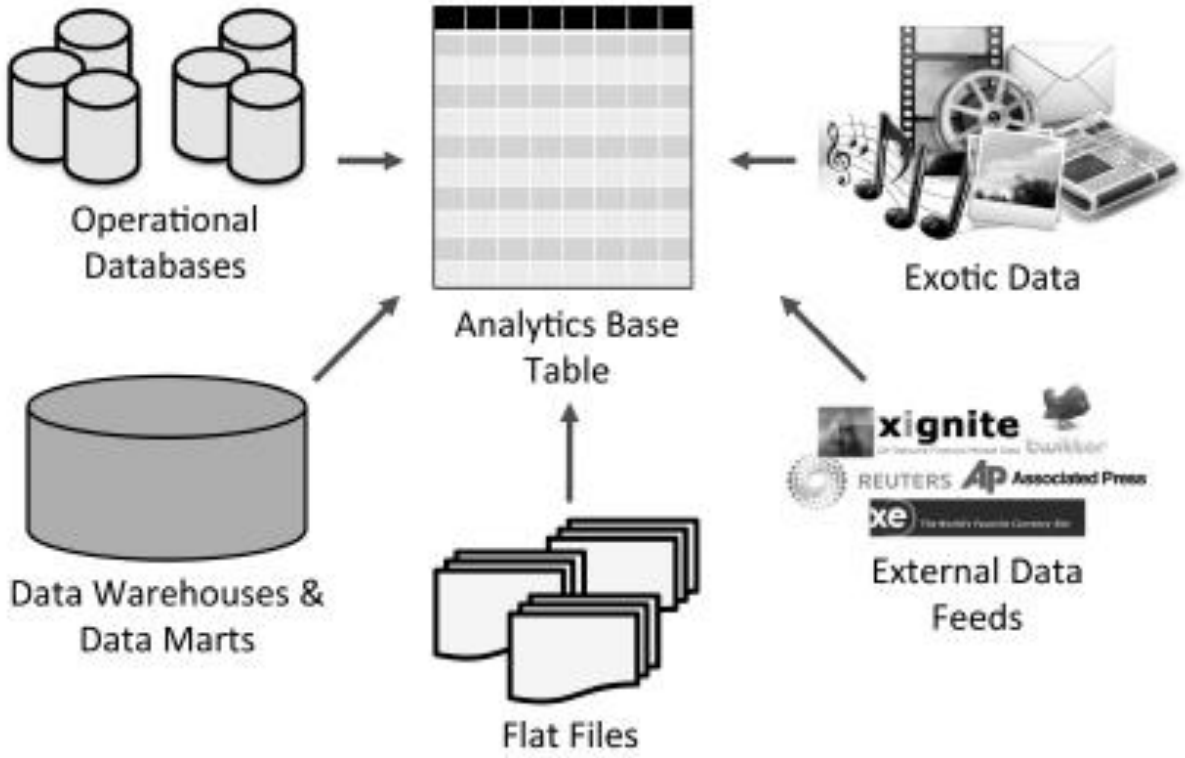


Figure: The different data sources typically combined to create an analytics base table.

ABT Design - Row

- Decide the **prediction subject**
  - The prediction subject defines the basic level at which predictions are made
  - 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.
  - Example
    - The prediction subject of the claim prediction model would be an insurance claim.
    - For the member prediction model, the prediction subject would be a member.

ABT Design - Features

- 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 define the features on these concepts.
- Make a hierarchical distinction between the actual features contained in an ABT and a set of domain concepts upon which features are based

Domain Concept, Descriptive Features, Target Feature

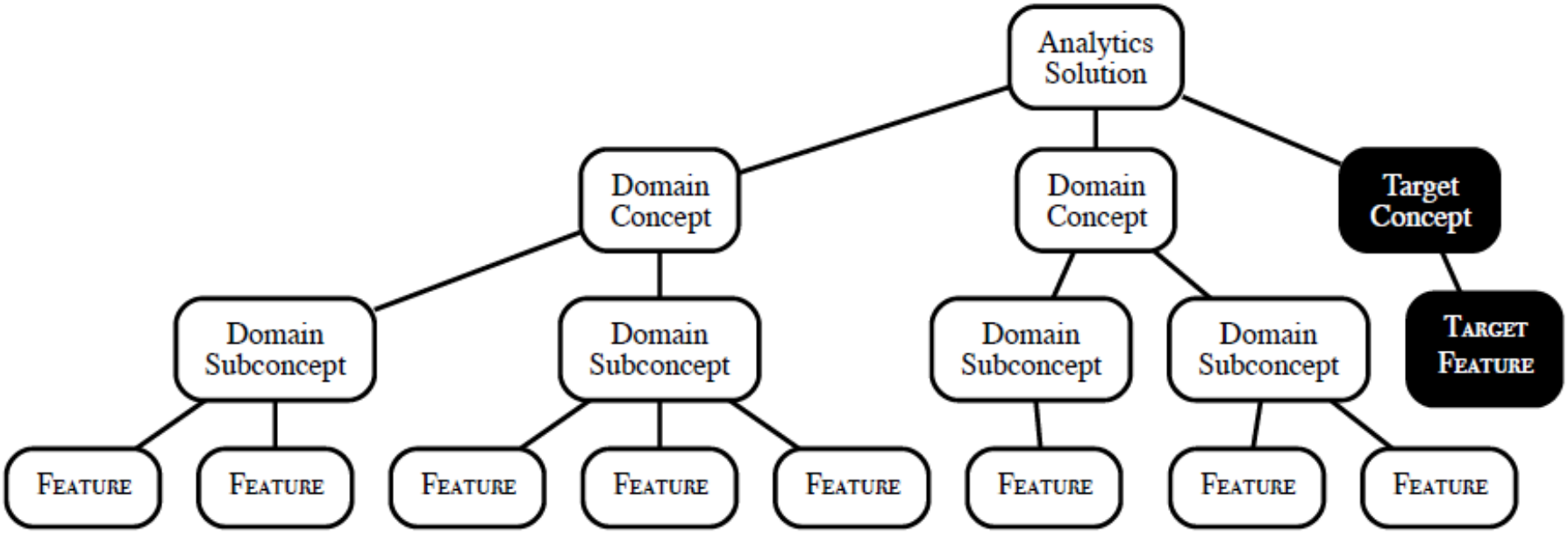


Figure: The hierarchical relationship between an analytics solution, domain concepts, and descriptive features.

ABT Design – General Domain Concepts

- There are a number of general domain concepts that are often useful:
  - **Prediction subject details**
    - Descriptive details of any aspect of the prediction subject
  - **Demographics**
    - Demographic features of users or customers
  - **Usage**
    - The frequency and recency with which customers or users have interacted with an organization
    - The monetary value of a customer's interactions with a service…
  - **Changes in usage**
    - Any changes in the frequency, recency, or monetary value of a customer's or user's interactions with an organization (E.g., example, has a cable TV sub- scriber changed packages in recent months?).

ABT Design – General Domain Concepts

- (Cont..)
  - **Special usage**
    - How often a user or customer used services that an organization considers special in some way in the recent past (E.g., has a customer called a customer complaints department in the last month?)
  - **Lifecycle phase**
    - The position of a customer or user in their lifecycle (E.g., is a customer a new customer, a loyal customer, or a lapsing customer?).
  - **Network links**
    - Links between an item and other related items (E.g.,, links between different customers or different products, or social network links between customers).

Motor Insurance Fraud – Domain Concepts

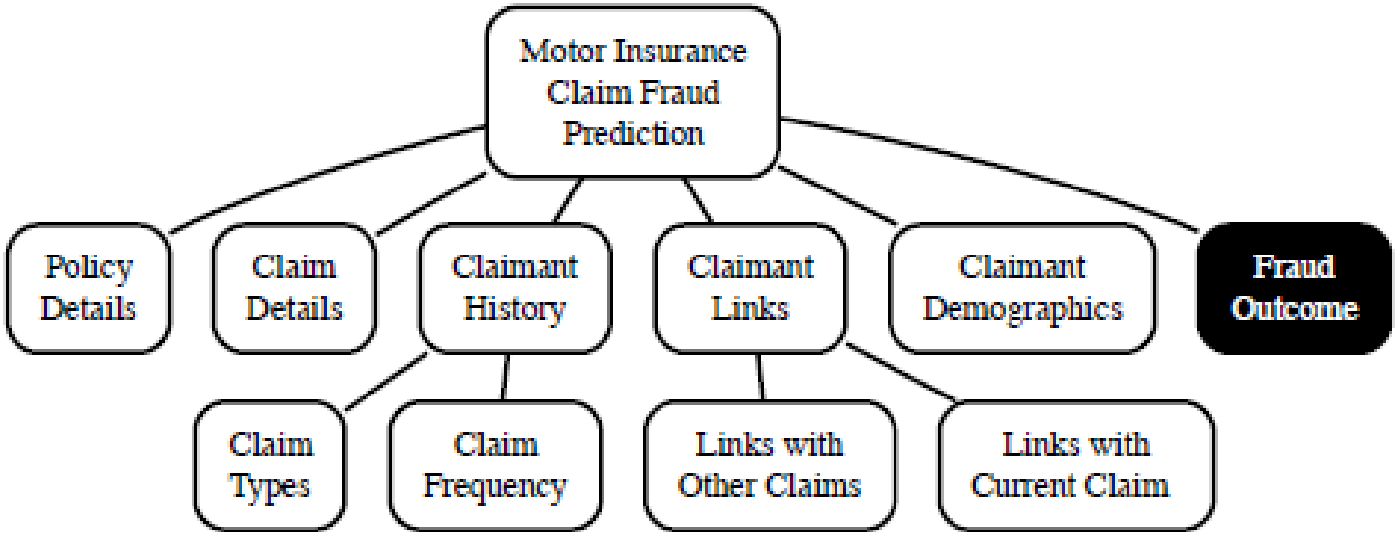


Figure: Example domain concepts for a motor insurance fraud claim prediction analytics solution.

Outline

- Converting Business Problems into Analytics Solutions
- Assessing Feasibility
- Designing the Analytics Base Table
- **Designing & Implementing Features**
  - Different Types of Data
  - Different Types of Features
  - Handling Time
  - Legal Issues
  - Implementing Features
  - Case Study: Motor Insurance Fraud
- Summary

Designing & Implementing Features

- Understanding and exploring the data sources related to each domain concept that are available within an organization is a fundamental component of feature design.
- Three key data considerations for designing features.
  - **Data availability**
  - **Timing**
  - **Longevity**

Data Availability

- We must have data available to implement any feature we would like to use.

Timing

- Data that will be used to define a feature must be available before the event around which we are trying to make predictions occurs.

Longevity of Feature

- There is the potential for features to go stale if something about the environment from which they are generated changes.
  - E.g., the borrower's salary as a descriptive feature. Salaries, however, change all the time based on inflation and other socioeconomic factors.
- One way to extend the longevity of a feature is to use a derived ratio instead of a raw feature
  - E.g., in the loan scenario a ratio between salary and requested loan amount

Different Types of Data

- **Numeric**: True numeric values that allow arithmetic operations (e.g., price, age) ,
- **Interval**: Values that allow ordering and subtraction, but do not allow other arithmetic operations (e.g., date, time) ,
- **Ordinal**: Values that allow ordering but do not permit arithmetic (e.g., size measured as small, medium, or large)
- **Categorical**: A finite set of values that cannot be ordered and allow no arithmetic (e.g., country, product type) ,
- **Binary**: A set of just two values (e.g., gender) ,
- **Textual**: Free-form, usually short, text data (e.g., name, address)

Different Types of Data

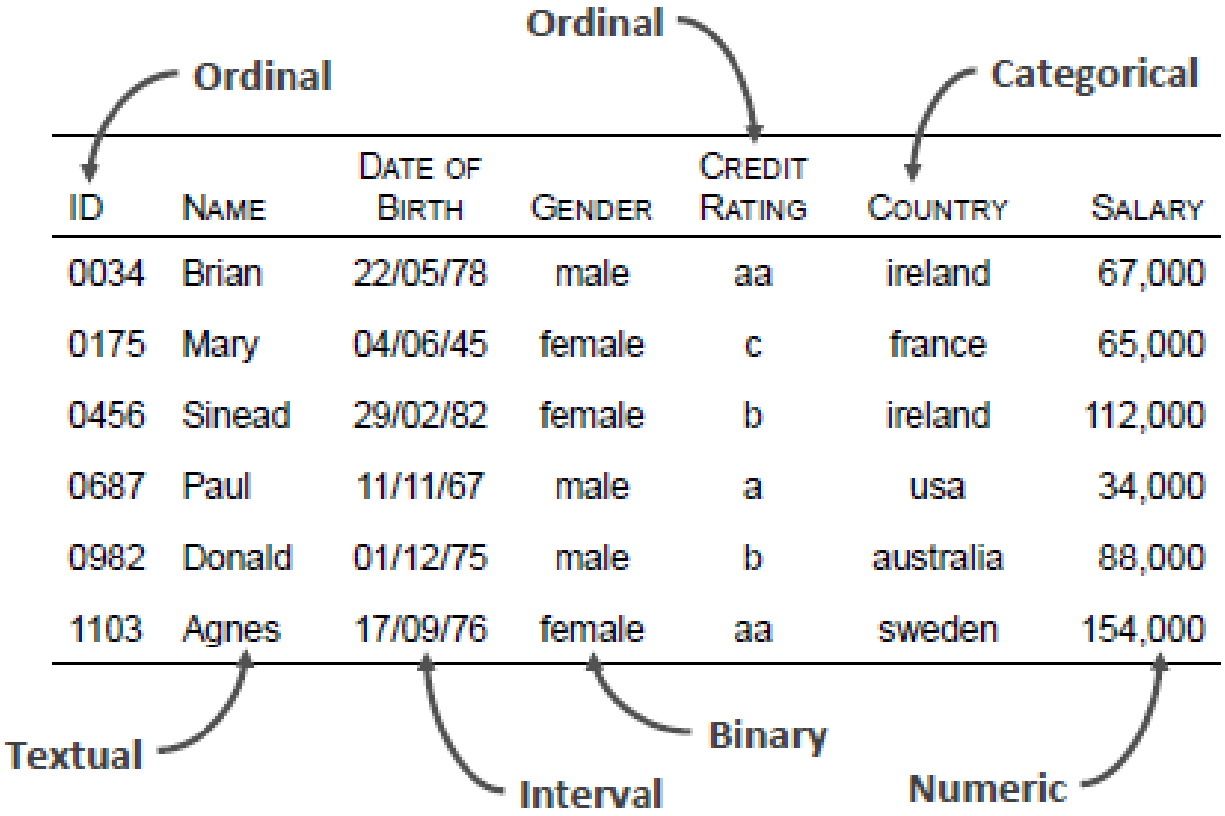


Figure: Sample descriptive feature data illustrating numeric, binary, ordinal, interval, categorical, and textual types.

Two Types of Data

- **Continuous**
  - encompassing the numeric and interval types),
- **Categorical**
  - encompassing the categorical, ordinal, binary, and textual types
  - a categorical feature can take as the **levels** of the feature or the **domain** of the feature.

Outline

- Converting Business Problems into Analytics Solutions
- Assessing Feasibility
- Designing the Analytics Base Table
- Designing & Implementing Features
  - Different Types of Data
  - ☞ **Different Types of Features**
  - Handling Time
  - Legal Issues
  - Implementing Features
  - Case Study: Motor Insurance Fraud
- Summary

Different Types of Features

- The features in an ABT can be of two types:
  - **raw features**
  - **derived features**
- There are no restrictions to the ways in which we can combine data to make derived features.
- Common derived feature types:
  - **Aggregates**
  - **Flags**
  - **Ratios**
  - **Mappings**

Aggregate Feature

- Aggregate features are usually defined as the count, sum, average, minimum, or maximum of the values within a group or a period
- E.g.,
  - the total number of insurance claims that a member of an insurance company has made over his or her lifetime
  - the average amount of money spent by a customer at an online retailer over periods of one, three, and six months

Flags Feature

- Flags are binary features that indicate the presence or absence of some characteristic within a dataset.
- For example, a flag indicating whether or not a bank account has ever been overdrawn

Ratio Feature

- Ratios are continuous features that capture the relationship between two or more raw data values.
- For example,
  - in a banking scenario, we might include a ratio between a loan applicant's salary and the amount for which they are requesting a loan rather than including these two values themselves.
  - In a mobile phone scenario, we might include three ratio features to indicate the mix between voice, data, and SMS services that a customer uses.

Mapping Feature

- Mappings are used to convert continuous features into categorical features and are often used to reduce the number of unique values that a model will have to deal with.
- For example, rather than using a continuous feature measuring salary, we might instead map the salary values to low, medium, and high levels to create a categorical feature.

Implementing Target Feature

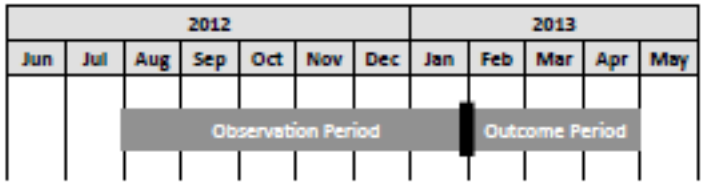
- Implementing the target feature for an ABT can demand significant effort.
  - For example,
- Just like descriptive features, target features are based on a domain concept, and we must determine what actual implementation is useful, feasible, and correct according to the specifics of the domain in question.

Handling Time

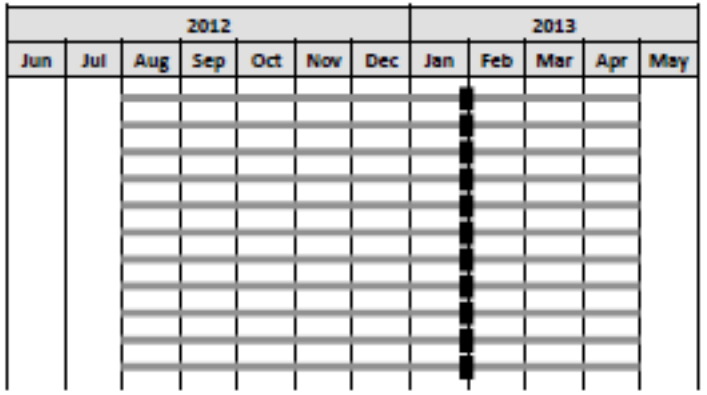
- Many of the predictive models that we build are **propensity models**, which inherently have a temporal element
- For **propensity modeling**, there are two key periods:
  - the **observation period**
  - the **outcome period**

Handling Time

- 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.

Handling Time

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

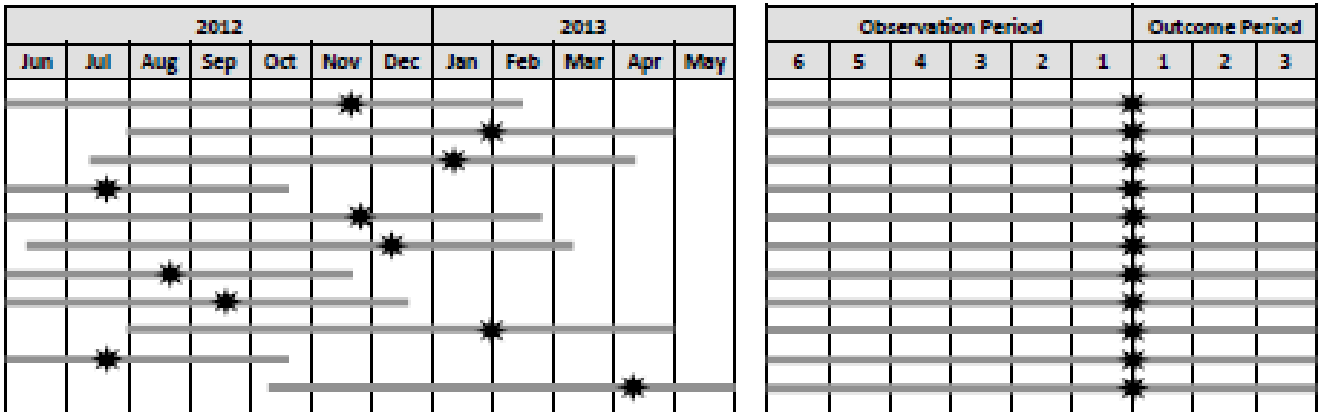


Figure: Observation and outcome periods defined by an event rather than by a fixed point in time (each line represents a prediction subject and stars signify events).

Handling Time

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

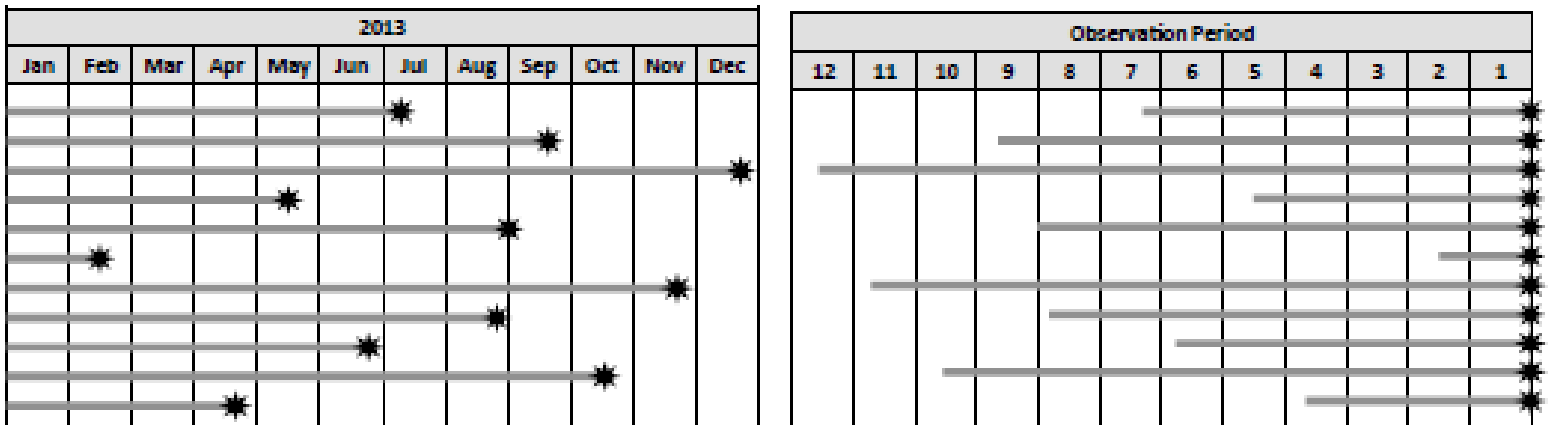


Figure: Modeling points in time for a scenario with no real outcome period (each line represents a customer, and stars signify events).

Handling Time

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

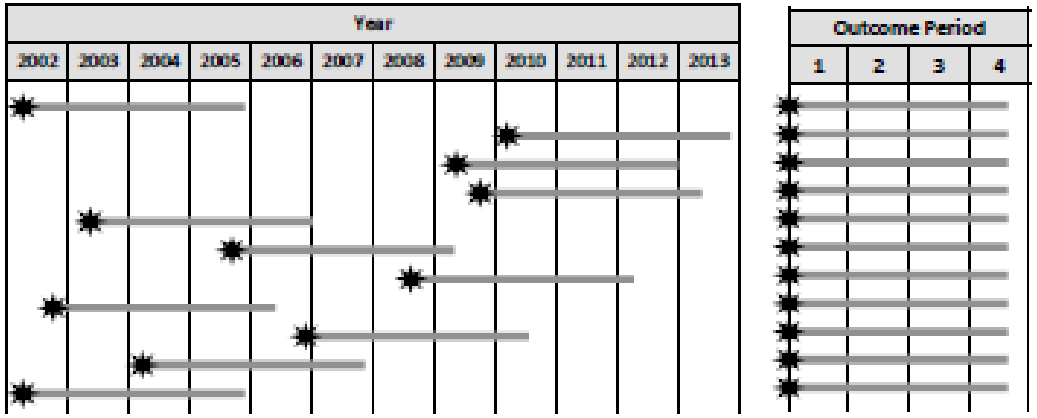


Figure: Modeling points in time for a scenario with no real observation period (each line represents a customer, and stars signify events).

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**

Legal Issues

- 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**

Outline

- Converting Business Problems into Analytics Solutions
- Assessing Feasibility
- Designing the Analytics Base Table
- Designing & Implementing Features
  - Different Types of Data
  - Different Types of Features
  - Handling Time
  - Legal Issues
  - ☞ **Implementing Features**
  - Case Study: Motor Insurance Fraud
- Summary

Implementing Features

- Implementation of the technical processes that are needed to extract, create, and aggregate the features into an ABT.

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?
- 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.

Motor Insurance Fraud - Domain Concepts

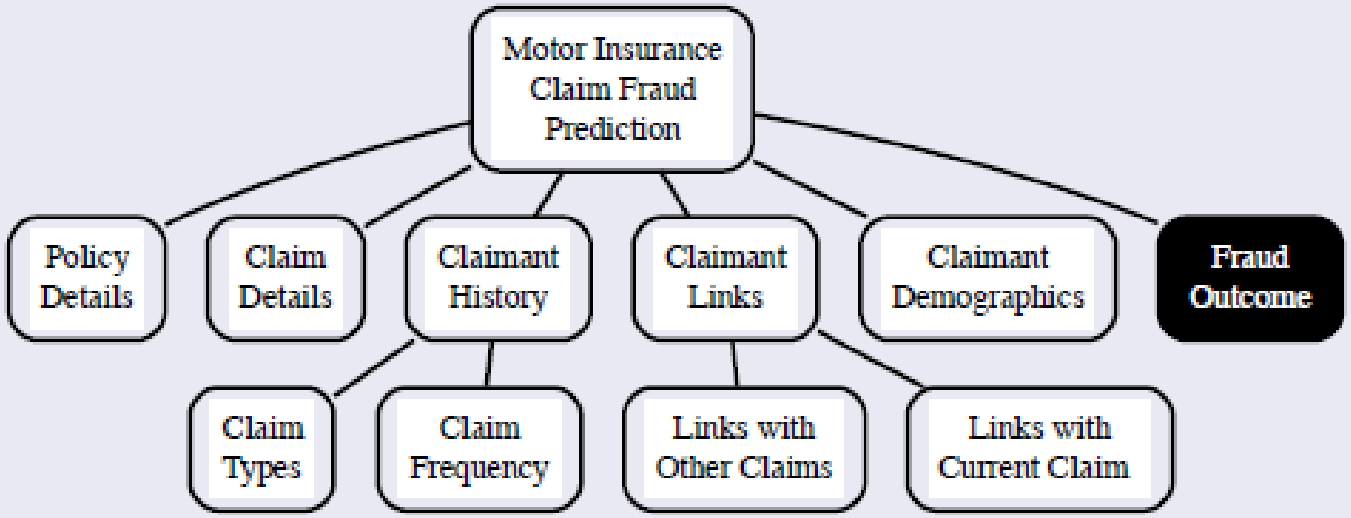


Figure: Example domain concepts for a motor insurance fraud prediction analytics solution.

Claim Frequency Domain Concept

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

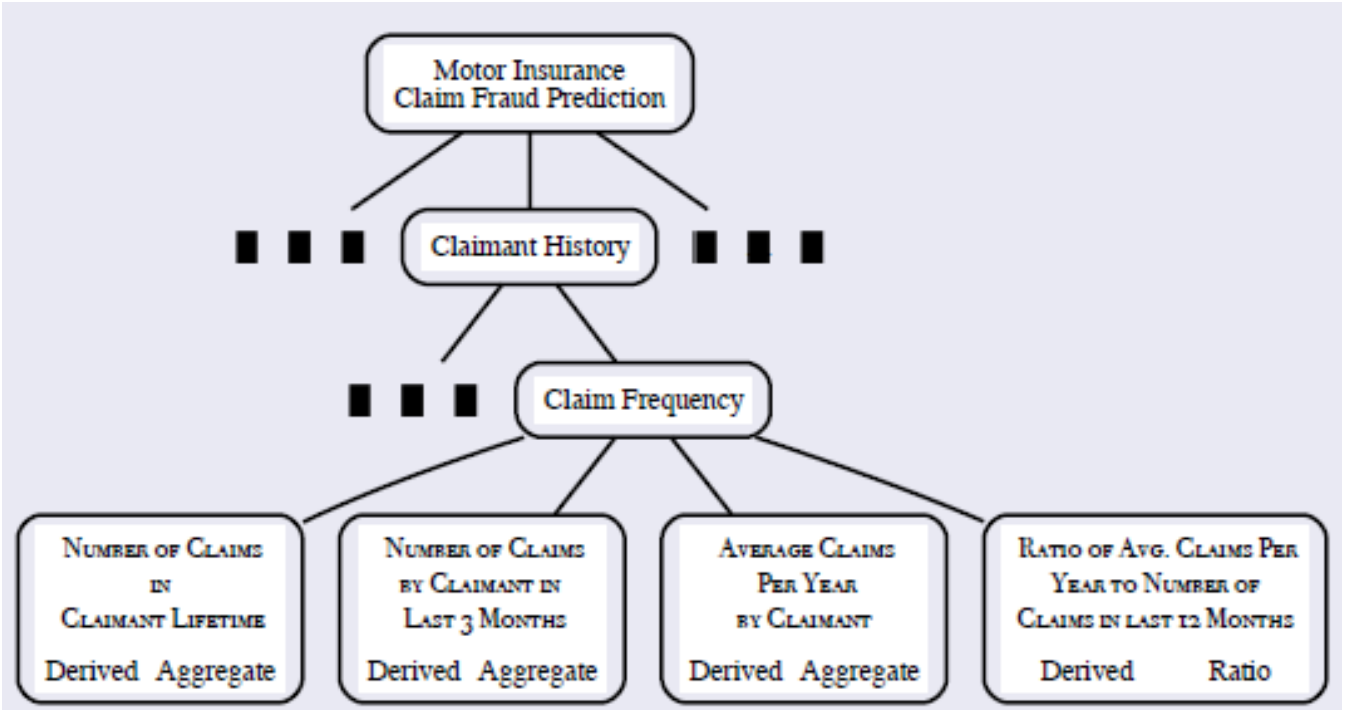


Figure: A subset of the domain concepts and related features for a motor insurance fraud prediction analytics solution.

Claim Types Domain Concept

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

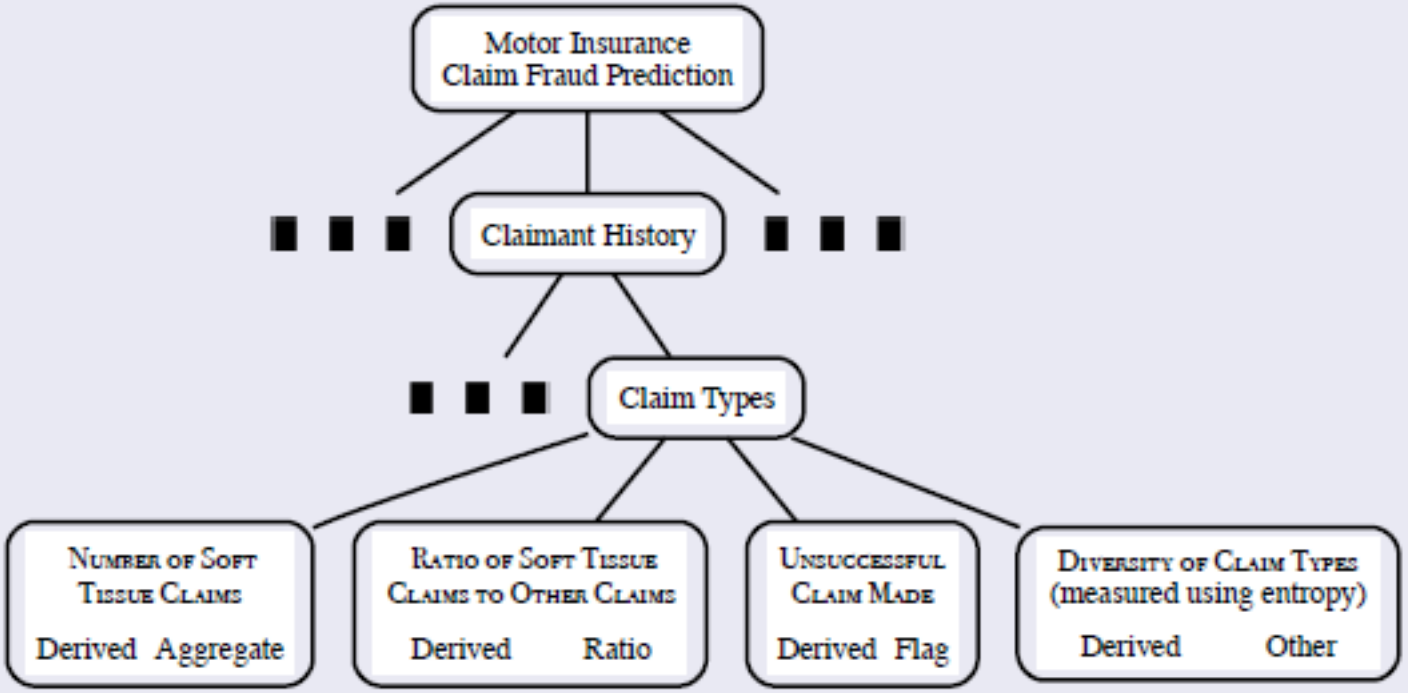


Figure: A subset of the domain concepts and related features for a motor insurance fraud prediction analytics solution.

Claim Details Domain Concept

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

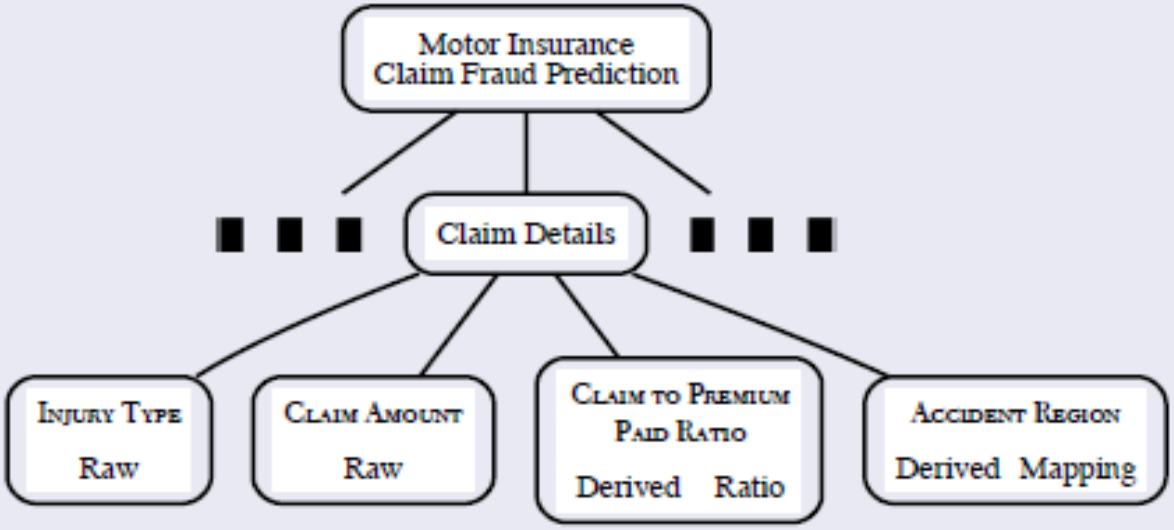


Figure: A subset of the domain concepts and related features for a motor insurance fraud prediction analytics solution.

ABT for 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
		⋮			⋮	

Outline

- Converting Business Problems into Analytics Solutions
- Assessing Feasibility
- Designing the Analytics Base Table
- Designing & Implementing Features
  - Different Types of Data
  - Different Types of Features
  - Handling Time
  - Legal Issues
  - Implementing Features
  - Case Study: Motor Insurance Fraud
- ☞ **Summary**

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

Summary (Cont.)

- 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.

Summary (Cont.)

- 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.

Summary (Cont.)

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

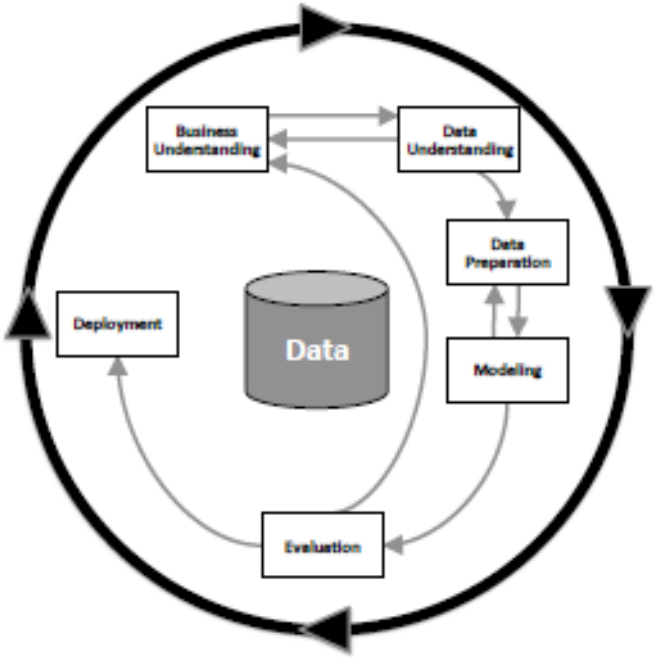


Figure: A diagram of the CRISP-DM process.

Summary (Cont.)

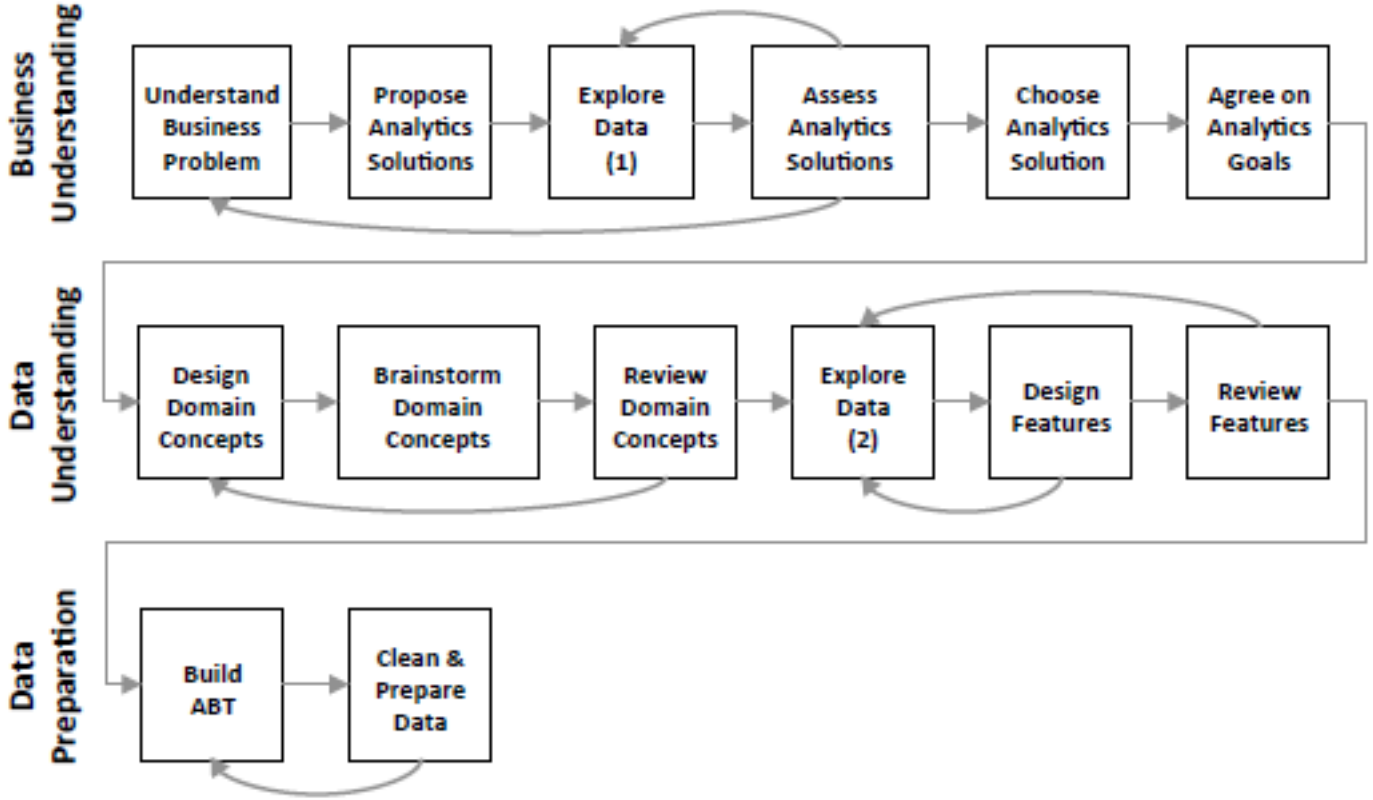


Figure: A summary of the tasks in the Business Understanding, Data Understanding, and Data Preparation phases of the CRISP-DM process.