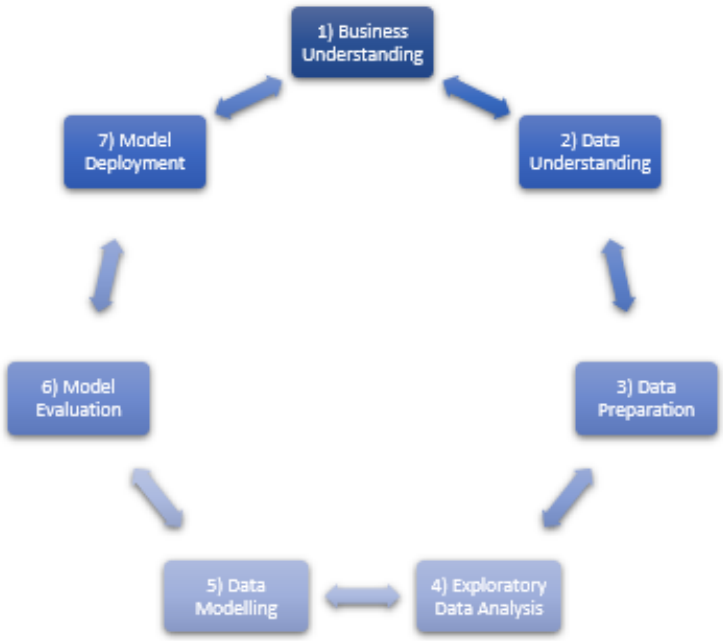
Predictive Model Building Framework

Within the Advanced Analytic space three basic predictive models are build

- Cross-sell
- Upsell
- Persistency

The Cross Industry Standard Process (CRISP) are an iterative framework on which every predictive model are developed. The steps in the framework are not linear data science steps. Starting with step one and then proceed to step two, however, from there, the naturally flow among the steps as necessary. Several small iterative steps are better than a few larger comprehensive phases.



1) Business Understanding

The starting point on building efficient propensity models is defining the business objectives, and it is necessary to map out the strategy before any model implementation begins. To build any propensity model, it is crucial to understand the what and the how the business is going to utilize these probabilities.

Example:

Business Unit	Propensity Model	Definition/Strategy	What	How
Risk	Funeral Propensity	Propensity to buy a funeral product in the next 12 months.	SLS wants to cross-sell on a group level (All Sanlam Brands) funeral products to clients	Through different channels: • Direct channel campaign • SFA & SBA through Unwrap campaigns

These goals will provide the basis for the next steps in the model building life cycle.

2) Data Understanding

The subsequent step is data understanding which entails a series of all the reachable data. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data. **It is also important to explore with business what type of information can influence the predictability. It is not just a case of throwing all the data available in the model, in some cases new data fields can be derived from the existing data which can add more value to the predictability of propensity model.**

From understanding what data is available it will answer the question whether the business need can be resolved. It might happen that the business need have to change because some data is not available.

Example:

To explain this practical lets look at our example above, starting of with the understanding of what type of information can influence the predictability

- Client demographic information: Age, income, gender, number of children dependents, marital status, etc...
- Client product portfolio information: What other products do the client have, total monthly premium amount, when last did the client purchase a product, how long have the client been with Sanlam, etc...
- Tradititional vs digital: Through which channel is most of the funeral products sold, and the client product purchase preference is known it can influence the predictability.
- Intermediary behavior: Are there certain intermediary information that trigger a funeral product purchase. If yes and we include this information in model - does this model still fulfill the business need?
- Are there certain events that can trigger a funeral product purchase?

Sanlam Analytical Data Sources

Within Sanlam there are **internal data** about the

- Centralized client demographics, product usage and their jockey intermediary information (also at a historical view)
- Detail product information (also at a historical view)
- Intermediary information
- Sentrix (Call centre data)

As well as **External data sources**

- Compuscan (Clients credit data)
- Home affair data

Data Source	Location*	Table/File name	Description
Client Data Hub (CDH)			

3) Preparation of Data

Next comes the data preparation stage. This consists of steps like:

- choosing the applicable data
- integrating the data by means of merging the data sets
- cleaning it
- treating the lacking values through either eliminating them or imputing them
- treating inaccurate data through eliminating them
- test for outliers the use of box plots and cope with them
- constructing new data
- derive new elements from present ones
- format the data into the preferred structure
- eliminate undesirable columns and features

Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

To explain this practical lets look at our example above, starting of with the understanding

- How many funeral products are sold per year? Are one year's of data enough to build the model - maybe it will require a longer period
- From above Q we might to look at a three year period, but there are several funeral product types under the funeral product main category.

4) Exploratory Data Analysis

This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model.

- Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs,
- Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps.
- Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

5) Propensity Model Building

Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

6) Model Evaluation

Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

7) Model Deployment

The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste.

For example, if data is no longer accumulated properly, you'll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.