General Predictive Modelling Principles

This page aims to provide a high-level overview of the general check points to consider when conducting the data management and predictive modelling phases of the Advanced Analytics data science value chain.

Data Management & Exploratory Data Analysis Check List

This section serves as a high-level checklist for ensuring a robust and hardened data management and exploratory data process analysis. This further validates that the data is free from error and ready for the predictive modelling process. The table below provides a list of components that aim to smoothen the data management process.

Component	Description
Identification of desired variables	After the business problem has been well thought through, the underlying attributes need to be identified and input as part of the data management process. These attributes, should satisfy the business problem and ultimately make sense for inclusion.
Data quality assessment	Once the various sources have been amalgamated, it is imperative to conduct a data quality assessment which encompasses (but not limited to) the following: • Nullability checks. – In some cases, variables with a large proportion of NULL values may not be worth including in the modelling process. In other cases, a NULL value may provide insight in itself. As an example, the credit bureau comprises a range of variables, many of which have NULL values to indicate the absence of such data for a specific client. Correct interpretation of the variable may lead you to conclude that this is, in itself, an insight - therefore making the variable worthy of inclusion. • *Outlier Removal - There could be cases within the data that skews the picture of certain measures. It might be best practice to omit these from the input predictive modelling dataset. This applies particularly to regression. • Assessment of required normalization within a specific attribute; e.g. Consider Gender that might encompass two categories to represent Male i.e., ‘Male’ and M. As such, this would require standardization to prevent skewness. • Assessment of aggregated measures in relation to the source data. – Essentially do the results make business sense for certain measures in your final data management dataset for the cohort that you are investigating. • Duplicate Check - Assess the level of duplicates after the amalgamation of various sources. In some cases you might have valid duplicates based on the modelling use case.
Feature Engineering	This is done in attempt to normalize or create variables that do not appear in source directly.
Cardinality of Variables	There could potentially exist attributes that have a wide range of elements. If possible, these should be reduced to a reasonable level so as to promote easier articulation.
Bivariate Analysis	This involves the study of the relationship between two attributes. This aids in understanding potential target leakages from the input dataset.
Target Variable Proportions	Assessing the target variable proportion is key in understanding whether or not the appropriate sampling techniques will need to be conducted in order to allow for “fair” predictive modelling. There is no desired benchmark (In a perfect world a 50/50 split would be favorable), however the desired adjustment level will need to be decided upon after a singular iteration of a model is run.
Consistency in terms of variable naming	Once the preceding steps have been conducted, ensuring that the final predictive modelling dataset has lowercase attribute names and be a 1-1 relationship between the source name. This ensures consistency between various data management platforms.

Target Variable and Class Definition	Just to ensure that all models are consistent in terms of the predictive output (especially for the API deployment scoring process) there would require a level of standardization in terms of the Target variable name and classes definitions (in the case of binary or multi-class predictive use-cases). The standardization principles fir regression and classification are highlighted below: **Regression:** Preferably if a variable was used directly from a source and used as a direct target input, the it would best practice to keep the name aligned. In other instances in which feature engineering was applied, the target variable can be user defined. **Classification:** For classification use cases it would be preferred to have the target variable called "target" and the underlying classes to be in numerical order (binary classification) & user defined (multi-classification) : *Binary Classification:* 1 - The class that you are interested in directly predicting. 0 - The opposing target class. *Multi-Classification:* *This can be user defined for multi-classification problems - to offer better insight especially if there are a variety of classes involved.*
Primary key identifier	Each unique row within your dataset should have a primary key either indicated by the client id, policy number or conjunction of the two. This allows for ease of traceability and integration.
Seamless integration of data modelling scripts	The code from the data management process should be easy to be wrapped into an automation process. This implies that the code should be free from noise and as clean as possible.

Predictive Analytics Check List

After the above-mentioned data management and EDA checks have been conducted, the next step would be to conduct the predictive modelling process. The table below provides a high-level check list to ensure a hardened predictive modelling process.

NB: Given that DataRobot is widely used for the predictive modelling calibration, the steps below are much more concise.

Component	**Description**
Desired feature list selection	Select the desired feature list from the input dataset. You could have a revised feature list from the input data ingested into DataRobot.
Target Variable and Class Definition	Ensure that the target variable and classes follow the standards are indicated in the EDA principles section.
Primary key masking	Ensure that the primary key is masked from the input feature list. This aids in the scoring process, so as to better identify a record /instance relating to a client. Some key examples would be the policy number, PID/pty_id and associated LID number/Member number of a policy holder.
Statistical performance evaluation	This refers to the evaluation of the predictive model using statistical measures. These will aim to assess whether a model is statistically sound. However, these results should be coupled with the business evaluation.
Business performance evaluation	From a business perspective, the predictive model needs to align with and solve the business problem. As such, there are a number of ways in which the performance can be measured from a business perspective: • Consider the spread of the actuals (of the target class) across the propensity spectrum, whilst using the **test** dataset. • Consider the spread of the actuals (of the target class) across the propensity spectrum, whilst using an unseen **batch** dataset (representative of the business at a point in time. – Ensuring a wider performance assessment.
*Predictive model tuning	This is centered on the hyper parameter tuning of an existing predictive model so as to improve the overall performance. – Within DataRobot, the out-of-the-box models are usually fairly predictive and limited fine-tuning is required.
Model Shareability	Ensure that the predictive model is shared with a member of the AA team for review. In addition to this, the model should be shared with the DataRobot admin (*with the Owner role applied*) to ensure a seamless deployment. **NB:** The admin email address is : *datarobot.adminapi@sanlam.co.za*