Current State Analysis - Data Capabilities

Evaluate the current BMO’s capabilities in terms of the data infrastructure. The data used to develop a machine learning model is of critical importance where high volume of data is managed due to the mining nature of this type of model.

The current state analysis is focused on the data sourcing, quality, availability, and maintenance for:

* Model Development
* Model Validation
* Model Implementation

1. **Inventory of existing tools and available data sources**

BMO contemplates internal and external data sources for modeling purposes. As a rule, developers should relied on internal data, as the most relevant and representative data of BMO’s portfolio. However, external data sources may be used when the internal data is insufficient for statistical purposes. The external sources may fulfilled the data rules and be representative data of BMO’s portfolio.

* 1. **Internal Data:**

The internal data is originated from different data source systems: TSYS, CCAPS, MBR, BASEL, MBA, CAD, among others, where most of the internal data is stored in the central repository,called**, Information Delivery Platform (IDP).**

IDP is a data lake repository where the Data stored follows three levels:

* Level 1: Raw data (Source Data) stored.
* Level 2: Conformed Data
* Level 3: Consumable Data. New Views (tables) are created. The views are created for users to get access to the data according to the model requirements.

These views are available to be used in other platforms /software, e.g. SAS, Dataiku, Excel, SQL.

However, as today, not all the internal data from the data sources is stored/ transferred in the IDP. For instance, IDP may have only a limited history from some of the sources. Similarly, not all variables from these sources are sent to IDP.

* 1. **External Data:**

For the current data structure, external data is divided into external data to IDP, and External data to the Bank.

* **External data to IDP** corresponds to internal data that were excluded from IDP, and data from other sources, e.g. macroeconomic data. For macroeconomic data, in particular, some of the History is stored in IDP (feed from Haver Analytics), while forecasting data, which is produced by the BMO Economics group.
* **External data to the Bank** refers to external data sources that are collected for a particular group for its internal analysis when the internal data does not suffice the model requirements. Different groups may have external data feeds set up at their level. Not all these feeds get to IDP. For instance, these feeds could come from SNL for financial data, Market data, among other sources.

1. **Data accessibility and availability and reporting capabilities.**

* For Model Development and Model Validation: Users can have access to data from IDP and other sources to develop and validate models on the software/platforms: SAS, Dataiku, Excel etc. Therefore, the data is not necessarily all in one place before model development.
* For Model Implementation and Reporting: There are specialized implementation platforms according to the type of model: Stress testing, IFRS9, Adjudication, Basel, among others. These platforms pull data from IDP for the most part although the data feeds are supplemented with non IDP sources when there are missing data in IDP from a quality or history perspective.

Usually only recent data (latest snapshot) is used for implementation purposes, where most feeds are monthly, where every implementation platform has its own data feed, monitoring, etc.

Additionally, there is a data repository in IDP called **the Common Data Model (CDM)** which has been used mainly for reporting. CDM is reconciled and monitored every month. As a future capability, the analytical infrastructure will be sourcing from CDM and augmenting the data as appropriate.

1. **Review of current information/data governance organization and responsibilities**

The Model Risk Guidelines from July, 2018, determine that Data is governed by the Enterprise Data Governance & Analytics Group and outlined in the Enterprise Data Governance Operating Directive. The responsible groups for data manipulation are presented below.

**3.1 For internal raw data**

Raw data generated in daily operations by Operating Groups is owned by the Operating Groups. Operating Groups are part of Enterprise Data Governance Operating Directive, and they are responsible for the accuracy, integrity and completeness of their data.

On the other hand, the Model Developers (as users of the data) are responsible for raising data quality issues through Data Issue Management Process of the Enterprise Data Governance Operating Directive, where data issues are prioritized and addressed.

**3.2 For data extraction and manipulation**

The Modelling team is responsible for the following data requirements specification tasks (except for specific Model areas where another group has explicit responsibility):

* Defining the data requirements
* Identifying the data sources when appropriate
* Documenting the data extraction code

The Model Developers should engage relevant stakeholders, including MV, as appropriate, on critical decisions concerning the choice of data, treatment of data, selection of variables and risk drivers subject to the stipulations for preserving MV independence during early engagement in Model Purpose and Business context.

The Model Developer should document the weaknesses and limitations of the data and analyze their impact on the overall Model. Significant data weaknesses or limitations that affect the Model’s effectiveness should be clearly documented and communicated to the Model Owner and Users and other stakeholders, including MV.

**3.3. For External data**

When the Model Developer decides to use external data or proxies, the Model Owner, Developer and/or any other group designated by the Model Owner or Operating Group, is responsible to ensure the accuracy, integrity, relevance and representativeness of such data. The Model Developer must document any assumptions associated with the data sample and the rationale for their selection as well as the source and ownership of the data.

1. **Perceived and potential data quality issues**

* The data sourcing for modeling is not necessarily all in one place. Most of the data is retrieved from IDP, however due to limited history or limited variables sent to IDP, users retrieve data from the original data sources. Without the readily accessibility to data using a central repository, users will spend an important part of their time compiling datasets rather than analyzing data.
* There is a big initiative in place to bring all data required for model development, model monitoring and model implementation to IDP. This will be both an augmentation and a curation exercise where the goal is to have all models sourcing from the same data, in a productionalized manner with all control and checks following the IDP framework.
* Regarding Data integration and access, individual data sources need more accurate definitions and common identifiers and elements among data marts need to be documented. For instance, the data dictionary does not provide variable description and valid values for all the variables presented, which increase the dependency on the data owners to understand the data.
* Regarding tools, Dataiku will be used for Development and Validation where applicable but it’s not the only tool. RCMD is still mainly a SAS shop and 99% of BMO’s models are developed in SAS. On the other hand, the implementation process will continue on the implementation platforms. Dataiku doesn’t implement models. It has an API creation functionality for example but it’s not an implementation platform.
* For emerging data, the associated data governance has not been established yet to our group knowledge. The data governance group might have starting thoughts on this although this topic

Please refer to the ***Risk Modeling Data Flows***file for a “futuristic” view of the data flows in the risk models.

**References**

Model Risk Guidelines, July, 2018.