Data Audit Report

A Data Audit report consists of -

1. Data Catalog
2. Data Quality Scorecard

**Data Catalog**

This catalog documents the current understanding of the data source, communicates to stakeholders the sources to be used and some basic facts about them, and helps identify potential mismatches, concerns, or clarify misunderstandings.

Another simple method of identifying potential challenges is to clearly diagram the data pipeline used to build the model, showing where all data is from and how it is transformed.

Th common information a Data Catalog contains –

* Major data cleaning operations performed
* Records dropped, either as an actual count or as a percentage
* Major issues found with the data, for example, “duplicate records found and dropped”
* Assumptions made at the time, for example, “data was extracted for US only, other countries are assumed to be similar”

Example –

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Contents | Duration | Quantity | Comments |
| Data source  #1: data lake | Clickstream  data | Jan 2018– Jan 2019 | 1.6M | User IP address only;  user name not known |
| Data source  #2: data lake | Order history | June 1 2016–  Oct 3 2018 | 55k orders | Format stored in  changed on Jan 1 2018  Final order only  (not change history)  Orders with errors  are deleted |
| Sensor data | Readings from factory sensors. Streaming data is batched and stored | 90 days history  retention only | 50/sec; 5k/sec  expected | Data cleaning unknown;  is perceived outlier data  being dropped? |

**Data Quality Scorecard**

The success of a Data project depends on the signal inherent in and extracted from the data. The Data Quality scorecard highlights areas that are frequently problematic in Data analysis and data science projects, and that can easily mislead the project—missing a signal that exists in the data

or believing a signal exists where there is none—if not identified and addressed.

Example -

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Category | Example | Issue for project? (Y/N) | Status (Red/ Yellow/ Green) | Comments: Applicability, Status, Mitigations |
| Input data precision | Test & production data have  same characteristics; outliers discarded for both model & production |  |  |  |
| Input data accuracy | Sensor values estimated to be +- 5% of actual |  |  |  |
| Data volumes &  duration | Model data only available for 3 months  (but business cycle is 1 year) 50% of source #3 data discarded |  |  |  |
| Data sources &  preprocessing validated | Data source #1 now undergoing  additional quality checks Data extract #2 discovered to be flawed; re-training required |  |  |  |
| Production vs model  data pipeline | Prod inferences will use separate  data source than model trained on |  |  |  |
| Data change over  time: processes considered | Upstream system changes logic  & meaning of its input to model |  |  |  |