KPMG

The Manager

Dear Sir/madam

Re: Report on Data Quality Assessment- Sprocket Central Pty Ltd

Looking at the datasets in relation to data validity, data uniqueness, data precision\accuracy, Reliability, timeliness and consistency the following are some of the data quality issues I came across

There is inconsistency in some of commonly used attributes for instance, the gender column instead of having three possible values that is, male, female or U for Unisex, values such as femal and F came up, this inconsistence degrades data consistency

The default column provides irrelevant information, it contains different character forms which are not related and makes no sense therefore, it should not be included in the dataset. While collecting data one should consider the relevance of the feature before including it.

Data completeness is also an issue here, some of the most essential attributes have missing values. Features such job industry category and job title have over 500 missing values. Incompleteness in data renders our data unreliable. To avoid this, you are advised to apply NOT NULL constraints for essential information in your data collection tools to enforce record entry

Date of birth for customer ID 34, year being 1843 is inaccurate figure. This wrong value interferes with data accuracy. This also affects data validity since age 175 is an invalid entry. Such age does not exist and will cause low accuracy in our model development.

From the task’s description, data should be for the last three months but after carefully looking into data, it covers data for twelve months, period between month of January 2017 to December 2017. Therefore, you should clarify what months or period you wish us to look into

Information concerning customers should be stored together to enhance uniqueness, for instance customer demographics and customer address should be stored together on the basis of customer id, this speeds the look -up process and minimizes redundancy for instance no need to have customer id in both sheets.

Labeling was an issue. For the deceased indicator, instead of using names such as N and Y prefer using Yes and No for easier data understanding. The column name ‘product\_sold\_first\_date’ do not bring out clear meaning of what the column is about. Therefore, you should clarify what this attribute represents

Regards

Bramuel Aswani