Position Paper 2-

Data quality for Predictive Modelling

Merrimack College

Data Exploration

Amol Gote

(Data science student)

June 2nd, 2020

*"Predictive Modeling is the process of using data and statistical methods to construct models, which can be used to predict values for unknown or future events." (www.microstrategy.com).* Data that is used in these models must be of high quality for the model to be successful in its business objective. This historical data which is collected can potentially have several data quality issues.

Firstly, the data should not contain duplicates records. For example, in a dataset same customer might exist with a different phone number or email, so as part of the data cleansing process, data needs to be de-duped. If the data is residing in the database, then some of the databases provide data quality services where de-duping can be done based on certain rules e.g. DQS Service from Microsoft SQL Server. The second potential data quality issue is data completeness or missing values, *"Data is considered "complete" when it fulfills expectations of comprehensiveness." (Sarfin 2019).* E.g. In case of patients dataset for certain dates, it contains first and last name and all relevant health record but is missing address and phone number then it is considered 100% complete if the predictive model does not factor in geographical location, if it considers the location then it is not complete. To address the completeness issue a summary table is created with each attribute in the dataset and percentage completeness. Once the percentage completeness has been identified, based on the attribute data type, appropriate imputation can be performed, if the data is numeric the missing value can be filled in with overall average dataset value. Apart from mean value substitution, median and mode can also be applied, this method is quick and can be easily done with in-built packages in R and Python, but the disadvantage with this method is it reduces the overall variance of the dataset if the applied to a larger set of variables. For numeric categorical variables, one of the options is to exclude the data from the dataset, another option is to make an educated guess e.g. in the survey response dataset for ratings 1 to 5 and user has responded all ratings with 4 then for missing value 4 rating value can be applied. Linear regression is an alternate method to treat missing values, in this approach predictive model is built based on the complete data and then the model is used to predict missing value. Theoretically, this approach provides a good estimate of missing value, but because it is predication based on the existing value it could reduce standard error. Hot-Deck imputation is another way of treating missing value, in this approach it is replaced with the random value from that column.

For maintaining data quality, outliers need to be identified, then flagged or removed. If the data set is small, then the data visualization technique of plotting the data on a scatter plot or box plot helps in identifying the outliers. In case if the data set is large, the mean and standard deviation is calculated for a variable, then all records with +-3 standard deviation are flagged. Data bias is another factor that can influence the predictive model results, to avoid that, skewness is the dataset is analyzed, this can be done with a histogram that shows the frequency distribution for a variable. To avoid bias, ensure that data is not over or underpopulated on specific variables and data distribution follows the standard normal curve. Data relevancy also impacts the predictive model results, data based on which the model is built must be directly related to business objective. If the business analyst or data-steward with business knowledge is involved in data quality checks, then the relevancy issue could be resolved easily. Apart from this data could have syntax errors like date getting stored in different formats like YYY-MM-DD or DD/MM/YYYY, currency columns containing currency signs like $, these syntactical errors can be corrected through code during data cleansing or data profiling.

The principle of *"Garbage in garbage out"* is a reminder that data quality is of prime significance for successful predictive models.

# Bibliography

Redman, T. (2018, April 02). *If Your Data Is Bad, Your Machine Learning Tools Are Useless.*

Retrieved from Harvard business review: <https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless>

*Definition: What is predictive modeling?* Retrieved from MicroStrategy website:

<https://www.microstrategy.com/us/resources/introductory-guides/predictive-modeling-the-only-guide-you-need>

*Data Matching.* Retrieved from Microsoft docs: [https://docs.microsoft.com/en-us/sql/data-](https://docs.microsoft.com/en-us/sql/data-quality-services/data-matching?view=sql-server-ver15)

[quality-services/data-matching?view=sql-server-ver15](https://docs.microsoft.com/en-us/sql/data-quality-services/data-matching?view=sql-server-ver15)

Sarfin, R (2019, Aug 7). *Data Quality Dimensions: How Do You Measure Up? (+*

*Downloadable Scorecard).* Retrieved from syncsort: [https://blog.syncsort.com/](https://blog.syncsort.com/2019/08/data-quality/data-quality-dimensions-measure/)

[2019/08/data-quality/data-quality-dimensions-measure/](https://blog.syncsort.com/2019/08/data-quality/data-quality-dimensions-measure/)

Swalin, A. (2018, Jan 30). How to Handle Missing Data. Retrieved from towards data science:

<https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4>