Joseph Annand

DSE5004

Position Paper #2

### Introduction

Data quality is essential in every data project, including predictive modeling, as poor data quality leads to poor analysis and misinformed business decisions. Lack of data quality has financial repercussions in millions of dollars for companies, thus it is important that data analysts and scientists understand how to evaluate data quality and remedy poor data quality (What is Data Quality?). There are several metrics used to measure the quality of the data being used for a given project. Each metric has its own impact on predictive modeling and business goals, but they can all be evaluated using a mix of visualization, statistical tests, and reason. If properly identified, issues across these metrics can be resolved before training predictive models and using those models to inform business decisions.

### Validity

Validity is the metric that measures how well the data matches required formats, such as metadata, data types, and ranges (What is Data Quality?). Data values must be formatted a specific way in order to meet the needs of the business or the model. For example, in a machine learning project that predicted the likelihood of a customer canceling their booking at a hotel, some features were formatted in the wrong data type for the purpose of the model; the numbers of children and adults for each booking were numeric when they needed to be factors that were then one-hot encoded for the neural network. In this specific instance, it was simple enough to convert the whole column of values into factors; however, problems could arise if some of the values were entered as words instead of numbers. Bar charts and pie charts can be used to see the proportions of different data types for each variable in the data set. Once identified, the values of the incorrect data type can be found and converted to the appropriate data type through basic logic.

### Outliers

Outliers are data points whose values are an abnormal distance from the rest of the population (NIST/SEMATECH e-Handbook of Statistical Methods, 2012). Models that are trained on data with outliers can be misled and produce inaccurate results. In my own work as a hemostasis scientist, outliers are not uncommon in laboratory experiments as several events can occur that produce an abnormal result when measuring blood clotting. When plotting the data and fitting a weighted linear regression, outliers become obvious as they will be far removed from the other points on the plot, and the linear fit will be unusually pulled in the direction of the outlier. Scatterplots are not the only way to determine outliers, though; box plots are also an effective way to see how removed a value is from the rest of the data (Tatanaki, et. al., 2024) and statistical tests, like Grubb’s test, can be used to determine if the value is an outlier for a data set with an approximately normal distribution (NIST/SEMATECH e-Handbook of Statistical Methods, 2012).

### Missing Values

Missing values can be detrimental to a predictive modeling project, as it may lead to biased results and misleading analysis (What is Data Quality?). There are different types of missing data, such as missing completely at random, missing at random, and missing not at random; each type of missing data may affect the analysis differently. Missing at random means that the probability of a data point is dependent on the values of other variables in the dataset (Kelta, 2023). The dataset can be filtered to find all observations with missing data points, and bar charts for each variable in the data set can be created to show the frequency of each value for each variable to find the reason for the data being missing. Once the reason for the data being missing is identified, the missing value can be imputed using the center of distribution statistics (mean, median, or mode), forward-fill or back-fill strategies, extrapolation, or k-nearest neighbor imputation (Ogunbiyi, 2022). The appropriate strategy may be dependent on the proportion of missing values, data type and the distribution of the data, and other knowledge specific to the project. Missing not at random means that the probability of the data point being missing is related to the missing value itself (Gupta, 2024). This means there is nothing within the data set itself to explain why the value is missing, making it more challenging to appropriately replace the value without negatively affecting the reliability of the data. The entire observation could be deleted if losing the missing data points will not affect the integrity of the data set. If deletion is too risky of an option, imputation using the information from the non-missing values of the variable to which the missing data belongs is also an option.

### Duplicate Values

Uniqueness is another metric of data quality that if not checked can result in poor results and inaccurate analysis. Having duplicate copies of the same observations in the data set will diminish the accuracy of the data set in reflecting the population that is being analyzed and throw off any models being trained by this data. In a project for an earlier course, the goal was to determine the most profitable city or country to work in for a data scientist based on salary and cost of living data. Two data sets were made available that included observations for countries and cities; some observations were unique to each data set, but some countries and cities appeared in both sets. Without using a proper join function to combine the two data sets, the joined data frame might have contained duplicate countries and cities, which would have disrupted the analysis. Bar charts and pie charts can be utilized to identify duplicate entries in the data set, and from there, the duplicate entries can be removed to preserve the integrity of the data (Tatanaki, et. al., 2024).

### Consistency

Another challenge in data quality that specifically arises when multiple sources of data are joined together is consistency. Data from multiple sources should be checked to ensure that the two sources reflect the same trends and relationships and thus align in the information they tell the analyst. When data sets are not consistent, this could mean there are errors within one or both sets, or the data sets are not appropriately related for the purpose of the project. Consistency can be checked using histograms or statistics like mean, median, and standard deviation to verify the distributions of data for each variable are similar in both data sets. Scatterplots and correlation coefficients can be used to verify relationships between variables are consistent in both data sets. If these checks show discrepancies between data sources, then each source can be checked for accuracy and other data quality checks. The analyst may want to revisit each source to ensure that all the data reflects observations from the same population.

### Conclusion

Data quality ensures that businesses are making properly informed decisions which are more likely to lead to increased profits and productivity. Each metric of data quality has unique influence on predictive modeling and analysis, and exploratory data analysis can illuminate issues with the data before the modeling phase of the project begins. Visualizations, such as bar charts, scatterplots, box plots, and histograms, can convey duplicates values, missing values, and outliers. Quantitative methods, like statistical tests, can highlight issues in data quality, while statistics may also provide a solution, such as in the case of missing values. Having a solid understanding of data quality, analysts and scientists can save time and money by working with accurate and reliable data.

### References

Gupta, M. (2024, February 28). *ML | Handling Missing Values*. GeeksforGeeks. Retrieved March 31, 2024, from https://www.geeksforgeeks.org/ml-handling-missing-values/

Kelta, Z. (2023, January). *Top Techniques to Handle Missing Values Every Data Scientist Should Know*. DataCamp. Retrieved March 31, 2024, from https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values

*NIST/SEMATECH e-Handbook of Statistical Methods*. (2012, April). NIST. Retrieved March 31, 2024, from https://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

Ogunbiyi, I. A. (2022, June 24). *How to Handle Missing Data in a Dataset*. freeCodeCamp. Retrieved March 31, 2024, from https://www.freecodecamp.org/news/how-to-handle-missing-data-in-a-dataset/

Tatanaki, H., Peters, M., & Wanpal, N. (2023, July 19). *Data Visualization for Data Quality: How to Spot and Fix Problems*. LinkedIn. Retrieved March 31, 2024, from https://www.linkedin.com/advice/0/how-can-you-use-data-visualization-identify-quality

*What Is Data Quality?* (n.d.). IBM. Retrieved March 31, 2024, from https://www.ibm.com/topics/data-quality