COMP47350 - Data Analytics

Data Quality Issues

&

Data Quality Issues

Question 2:

1.0 Data Quality Plan

1.1 Markdown of Features & Potential Handling Strategy:

Feature	Data Quality Issue	Handling Strategies
ShipsFromState	Missing Values (41%)	Do nothing
ShipsFromCountry	Missing Values (37%)	Do nothing
ConditionNotes	Missing Values (46%)	Deletion
SellerFeedbackRating	Outlier (Low)	Do nothing
SellerFeedbackCount	Outlier (High)	Do nothing
ShippingTime_minHours	Outlier (High)	Do nothing
ShippingTime_maxHours	Outlier (High)	Do nothing
ListingPrice	Outlier (High)	Do nothing
ShippingPrice	Outlier (High)	Do nothing

1.2 Analysis of Data Quality Plan Strategies:

1.2.1 Irregular Cardinality

Within this dataset, there does not appear to be any irregular cardinality. For this reason, no action need be taken with regard to this Data Quality Issue.

1.2.2 Missing Values

There are a significant amount of Missing values across the three above-mentioned features. The code in the attached notebook related to null values outlines the number of null rows. The are various approaches with regard to missing values, such as; imputation, complete case analysis, the dropping of the feature, or to simply do nothing. Each of these approaches should be based on the percentage of missing values.

In the case of imputation, this would involve the replacement of null values with the most plausible estimate, such as the average, median, or mode. However in this case, all three features fall above the recommended threshold for implementing imputation, of 30%. The risk of implementing imputation on such a high percentage of missing value could negatively impact the predictability of the data.

As the code indicates, a potential strategy to fill missing country data, should the state or province information be available is suggested. However, after analysis of the code, there appear to be no occurrences where the state or province's value are not null, and the country value is null.

As the values of the features *ShipFromState* and *ShipsFromCountry* represent a significant geographical location, the deletion of this feature is not recommended. The deletion of the rows where the features values are null is not advised either. The potential loss of data from the other features related to these rows would significantly have a negative impact on the predicatability of the data. Therefore, for the above-mentioned two features the recommended strategy is to do nothing.

In the case of *ConditionNotes*, the data represents inputted additional information about the product, delivery method of the product, the condition of the product, and additional information written by the seller which may make reference to the sellers reliability and rating. Similarly, it would not be advisable to implement imputation on this data, as it is not practical given this features type. In cases where the value of the feature is not null, the values do not provide any information not already known, such as perhaps information which could help fill null values in the *ShipFromState* and *ShipsFromCountry* features. Similarly, the values of the feature *ConditionnNotes* make reference to Shipping Price, Seller rating, estimated Shipping time, and locations from where the product will be shipped. These are values which are already represented more accruately by other features in the data. Deletion of null rows for this feature would too have a negative impact of other more valuable data that the rows contain across the other features. Given the above information, the proximity of the feature's amount of null values to the recommended threshold of 50% null values in a feature, and the unlikelihood of the feature impacting Continuous or Categorical statistics, it is recommended that this feature be deleted.

1.2.3 Outliers

There are a significant number of outliers within this dataset. They are noted in the Data Quality Plan above. All except *SellerFeedbackRating* refer to a high outlier presence. The first goal is to establish whether these outliers are valid or invalid due to errors in the data.

1.2.3.1 SellerFeedbackRating

From the data, it should be noted that two main sellers, one with the ID *1207135739271432339* and the other with ID *-5203594887869139290* are examples of the outliers present. Both have irregularly high shipping times, with maximum shipping time for some products able to take 1008. Perhaps this information could attribute to some of the lower than expected average. The data suggests that ratings in this data should not fall below about 85. Given the type of feature, it is likely that some sellers could receive a low rating for certain shipping, pricing, or reliability issues. For this reason, and as it is possible to receive low ratings or ratings of 0, it is fair to assume that these outliers are valid and not the result of human error. For that reason nothing should be done to clamp them.

1.2.3.2 SellerFeedbackCount

The data suggests to large outliers in relation the count of feedback submitted to sellers. It is represented largely by two sellers. One seller is a featured merchant, whose product at a certain time allowed that seller to be marked as a winner. As this feature refers to feedback, it is possible that the outliers here have an effiective process which encourages feedback. For this reason, it can be assumed that this data is valid, and nothing should be done.

1.2.3.3 ShippingTime_minHours

This data may suggest that some sellers are setting default minimum shipping times as one seller has most products set at a minimum of 672 hours. Alternatively, perhaps some of these sellers frequently must ship internationally and therefore need to allow for more time. This particular seller does no provide any information about Condition notes, Country, or State/Province howevers. It should however be noted that this is the same seller, *1207135739271432339,* who received low feedback ratings that marked that seller as an outlier in that feature. The same seller occurs as an outlier in the following feature, which will now be discussed.

1.2.3.4 ShippingTime_maxHours

Given the above data about this one seller, it is possible that this is the case for all present outliers. However, these outliers do not provide us with information which allow us to understand what it takes to be a winner, as these outliers are not winners. For these reasons, nothing should be done in respect to these two feature outliers

1.2.3.5 ListingPrice

As there are no negative outlier in relation to the sellers listing price, it is safe to assume that these listing prices are valid and not the result invalid input. Due to the volume of outlier, it is not recommended that anything should be done to clamp these outliers. Doing so, could dramatically alter relevant continuous statisical information.

1.2.3.5 ShippingPrice

A significant amount of data about were these outliers are being shipped from is missing. Therefore it is difficult to assess whether theses outliers are due international delivering costs. *SellerFeedbackRating* data suggests that although cost may be high, customer satifaction appears to be quite high. Perhaps the increased prices are due to a higher quality of product being delivered, however *ConditionNotes* are null and provide no information to confirm or deny this. Similar to the above reason, changing or clamping these outliers could significantly alter statistical data, and therefore it is advised that nothing be done.

1.3 Conclusion

The task for this dataset is to look for indications across the features in the data that will result in a seller becoming selected a winner, denoted by a 1 in the *IsWinner* feature. The above action and strategies taken I believe should aid and not impede that goal.