DATA QUALITY PLAN

HOMEWORK 1

Giuseppe Benanti –ID 07266120

Feature	Data Quality Issue	Potential Handling Strategy
ListingPrice	Outliers (high)	Do nothing
ShippingPrice	Outliers (high)	Do nothing
ShippingTime_minHours	Outliers (high)	Do nothing
ShippingTime_maxHours	Outliers (high)	Do nothing
SellerFeedbackRating	Outliers (low)	Do nothing
SellerFeedbackCount	Outliers (high)	Do nothing
ConditionNotes	Missing Values (46.9%)	Drop feature
ShipsFromCountry	Missing Values (37.2%)	Drop feature
ShipsFromState	Missing Values (41.2%)	Drop feature
ListingCurrency	Cardinality = 1	Drop feature
ShippingCurrency	Cardinality = 1	Drop feature
ShippingTime_availtype	Cardinality = 1	Drop feature
SubCondition	Cardinality = 1	Drop feature
MarketplaceId	Cardinality = 1	Drop feature
ShipsDomestically	Cardinality = 1	Drop feature
TimeOfOfferChange	Cardinality = 548 (too high)	Binning or drop feature
IsWinner	Binary interpreted as numeric	Change values; 'Yes' for one and 'No' for zero
IsFeaturedMerchant	Binary interpreted as numeric	Change values; 'Yes' for one and 'No' for zero
IsFulfilledByAmazon	Binary interpreted as numeric	Change values; 'Yes' for one and 'No' for zero

NOTES

- Outliers in 'SellerFeedbackRating' are not dropped because probably not errors but real values of people not having a rating. Instead we could use a clamp transformation, choosing as a threshold 3 standard deviations below the median.

  Median = 95.0

  St.Dev. = 21.6

  Threshold = 95 – (21.6 * 3) = 30.2

  We choose the median and 3 st.dev. instead of 2 because the distribution is not normal but skewed to the left. In this case, the median will better represent the real centre of the distribution compared to the mean.

  But because we are not sure about this, we do nothing at the moment.

- We drop the features with missing values and not use any mean manipulation on them because of the nature of the features: for instance, Shipping Countries, it would not make any sense to assign to it the mode of the most occurring Country for other offers. The percentage of missing values is as well very high (above 30%), so using mean manipulation could change entirely the distribution of the features.

- After analysis of count of outliers, we decide to deal with them doing nothing; we do not drop the values because we don't believe they are errors, as they are counted in the order of hundreds over a dataset of 10000 entries and also because looking at the outliers they do not seem to be errors.

Giuseppe Benanti – 04/03/2017