Data Quality Report

1. Overview

This data quality report concerns the data in CustomerChurn-12751005.csv and is to used in conjunction with the Robert_Debhal_12751005_COMP47350_Homework1.ipynb Jupyter notebook.

Running the code in this notebook will generate the diagrams and desriptive statistics which are necessary to understand the initial findings discussed below.

2. Initial Findings

2.1 Data Types

The data types in CustomerChurn-12751005.csv needed to be adjusted for features of type bool and object. These were changed to type categorical to allow for ease of analysis.

2.2 Duplicate rows or columns and constant columns

The data contained no duplicate rows or columns or constant columns so no further adjustments needed to be made to clean the data set.

2.3 Descriptive Statistics

Descriptive statististics were calculated for both categorical and continuous features which gave some insight into the quality of the data.

For the continuous features, the min statistic highlighted some irregularities in the data. For example there were features with 0 as a minimum value where this did not make sense e.g. age and currentHandsetPrice. These statistics probably represent errors in the data collection or missing values.

For the categorical features, the count was very low for occupation and regionType, 255 and 508 out of 1000 respectively, indicating a large amount of missing data. Similarly, in the case of marriageStatus unknown was the most frequent category with 396 out of 1000.

2.4 Plots

Histograms and box plots were calculated for all continuous features, and these both showed the unusually high amount of 0 values in the data.

The histograms were largely skewed right, which made sense for some features e.g. handsetAge and lifeTime. For features such as age and currentHandsetPrice however, this suggested possible errors in the data, because a more normal distribution would be expected for these features.

The box plots for a large number of features had both a min and a 25th percentile of 0 whcih was very unusual for features like income, age and currentHandsetPrice. There were also a large number of features with outliers.

Bar plots were plotted for the categorical features, and these highlighted the irregular cardinalities of some features e.g. occupation and creditCard.

3. Conclusion

After performing an initial analysis of the plots and statistics for the data, it was clear that the features would need to be analysed in more detail to address the issues raised in the sectons above. This is done in the DataQualityPlan.pdf