14-20 March 2022

1. Impute missing values
   1. Replacing missing numeric values with mean of column they belong to
      1. COMPLETE – missing values are replaced with mean of column if they are numerical, other types are not of concern for our specific dataset as none are present
   2. Replacing missing values of other types – drop record?
2. Sign up date feature
   1. Format variable into just year
   2. Format variable into years since signing up
3. Convert to factor
   1. Convert char features
      1. COMPLETE – columns of type character are transformed into factor, and levels are printed following conversion
   2. Convert Boolean features
      1. COMPLETE – if a column of type numeric has 2 unique values, column is converted to factor from numeric
   3. Convert “Kidhome” and “Teenhome” features
      1. COMPLETE – now converts columns of type numeric if there are =< 3 unique values
4. Print graphs for all columns
   1. Print different graphs for each type of column
      1. COMPLETE – boxplots/histograms for numeric columns, bar graphs for factor
5. Remove special columns
   1. Remove columns that have no predictive power
      1. COMPLETE – if column has constant values across all rows, column is dropped
   2. Remove and identify ID column
      1. COMPLETE – if column has unique values equal to row count, column is considered an identifier and is dropped
6. Remove outliers
   1. Outliers in factor
      1. Drop levels that only appear in 5% of records?
   2. Outliers in numeric
      1. Drop records if they are outside of a threshold (std, var, etc.)
7. Normalization
   1. Normalize numerical features
      1. Check for normal distribution in histograms
8. New columns
   1. Total spendings column
      1. Identify top spenders/least spenders, campaign records
9. Minority class imbalance
   1. SMOTE
      1. Apply SMOTE and observe impact on accuracy, negative predictive value (lower than PPV)

21-27 March 2022

1. Sign up date feature
   1. Format variable into just year
      1. COMPLETE – each value in column is replaced by year, and are converted to factor along with other columns
   2. Format variable into years since signing up
      1. DEPRECATED – unnecessary as years are already converted to factor
2. Remove outliers
   1. Outliers in factor
      1. COMPLETE – drop rows that contain a factor value that appears in only 2% of rows in any given factor column
   2. Outliers in numeric
      1. COMPLETE – capping implemented for numeric columns, change value of outliers above 1.5\*IQR + 3rd quantile and below 1st quantile – 1.5\*IQR
         1. Formula can be changed to accommodate for dataset distribution
3. Normalization
   1. Min-max normalization
      1. COMPLETE – gather numeric columns’ index and apply normalization with “scale”
   2. Log scaling
      1. Investigate other scaling methods, impact is debatable
   3. Clipping
      1. Is clipping just capping?
4. New columns
   1. Total spendings column
      1. DEPRECATED – new column detection is difficult and impractical to implement, if columns present with collinearity, Pearson correlation scores can be used to drop them instead
5. Minority class imbalance
   1. SMOTE
      1. Apply SMOTE and observe impact on accuracy, negative predictive power (lower than PPV)
6. One-hot encoding
   1. Encode factors to dummy variables for modelling
      1. Required for neural network
      2. Ordering is important, as script prerequisite requires target column to be last
7. Algorithms
   1. Decision Tree
      1. COMPLETE – C5.0 algorithm already implemented
   2. Random Forest
      1. Script must run on Dr Fu’s computer for presentation, could be too heavy for a live demo
   3. Naïve Bayes
   4. Neural Network