Data Preparation, Analysis and Manipulation:

1. **Introduction:**

The standard data cleaning activity to prepare dataset for Machine Learning based prediction algorithm consists of primarily 2 steps as depicted below:

Raw Data

Technically Correct Data

Consistent Data

Type Checking and normalization

Fix and Impute

Fig-1: The 2-step data cleaning process (Ref: Jonge & Loo)

1. **Data Cleaning Approaches for the Customer Churn Dataset:** 
   1. Validity Checking and Data Imputation:

Typically, for any ML based prediction framework, there can be any number of validity rules used to clean the data, and these rules will depend upon the intended purpose or objective of the model. A summary of our data set is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. of Columns | No of Rows | No. of floating-point variables | No of Integer variables | Binary Logical | Categorical |
| 100 (including the predicted variable - Churn) | 100000 | 56 | 19 | 14 | 11 |

**Summary of Missing Values:**

|  |  |  |  |
| --- | --- | --- | --- |
| Total No of Columns having Missing values | Total count of missing values | % of Missing Values w.r.t the total dataset | No of columns that do not have any missing values |
| 26 | 157757 | 1.57% | 74 |

|  |  |  |
| --- | --- | --- |
| Fields | No. of missing values | % of total no of rows |
| numbcars | 49366 | 49.36% |
| lor | 30190 | 30.10% |
| income | 25436 | 25.43% |
| adults | 23019 | 23.01% |
| hnd\_webcap | 10188 | 10.18% |

**Top-5 columns with highest frequency of missing values**

* **The data validation & imputation steps that were done are as follows:**

**Type checking and consistency of the fields**:

No field with inconsistent data type was found. Values across the columns were consistent in the sense that there was no type mismatch between the values in the same column. However, we found almost 1.57% of the cell values as missing or ‘NA’. We imputed the missing values using various algorithm as explained in the next section.

**Data imputation for missing values:** Imputation is the process of estimating or deriving values for fields where data is missing. There is a vast body of literature available on various imputation techniques. For our data-preparation, we took scenario-based approaches for each field that had missing values:

* **Data Imputation for Numerical Variables:**

We adopted following approaches for numerical variables:

1. Mean: For numerical fields that had missing values distributed across a wide range, we imputed the missing values with the mean of the rest of the values for those fields.
2. Random sampling from the set of most frequently occurring values: For some of the numerical fields that had integer values and possibly representing some counts or frequencies we imputed the missing values through a random sampling from the set of values with highest frequencies.
3. Max Value: For some of the fields that had one value with overwhelmingly large frequency compared to others, we simply replaced the missing value with that particular value.

* **Data Imputation for Categorical Variables:**

Frequency Based Approach: For the categorical fields, we simply replaced the missing values with the one that is having the highest frequency. In our case, for most of the categorical variables, we had one or two values that were having overwhelmingly frequency compared to other values and hence the one with highest frequency was a rational choice.

* 1. Feature Engineering:

There’s a vast body of literature for feature engineering for datasets in order to apply ML algorithms.

Feature Engineering for Numerical Features include:

* Scaling:

Various scaling mechanism are in use for Machine Learning input dataset. The most commonly used ones are: Min-max normalization, Mean Normalization, Z-score normalization and Scaling to unit length.

**For our dataset, we used min-max normalization:**

* Encoding:

We used following encoding schemes for Categorical variables:

* + - * **Label Encoding:** Label encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. We used label encoding for binary logical fields that are ordinal.
      * **One-Hot encoding:** For non-ordinal categorical variables, we have used one-hot encoding that split one column with multiple categorical values into multiple columns with values ‘0’ and/or ‘1’

Reference:

1. Edwin de Jonge, Mark van der Loo: An introduction to data cleaning with R, Statistics Netherlands
2. <https://www.kdnuggets.com/2018/05/packt-tackle-common-data-cleaning-issues-r.html>
3. <https://blog.dominodatalab.com/manual-feature-engineering/>
4. <https://en.wikipedia.org/wiki/Feature_engineering>
5. https://en.wikipedia.org/wiki/Data\_transformation