ASSIGNMENT 2: Telco Customer Churn

Projects form an important part of the education of software engineers. They form an active method of teaching, as defined by Piaget, leading to a "training in self-discipline and voluntary effort", which is important to software engineering professionals. Two purposes served by these projects are: education in professional practice, and outcome-based assessment.

Data cleaning or data scrubbing is one of the most important steps previous to any data decision-making or modelling process. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

Data cleaning is the process that removes data that does not belong to the dataset or it is not useful for modelling purposes. Data transformation is the process of converting data from one format or structure into another format. Transformation processes can also be referred to as data wrangling, or data munging, transforming and mapping data from one "raw" data form into another format. **Essentially, real-world data is messy data and for model building: garbage data in is garbage out**.

This practical assignment belongs to Data Science Master at the UPC, any dataset for modelling purposes should include a first methodological step on **data preparation** about:

- Removing duplicate or irrelevant observations
- Fix structural errors (usually coding errors, trailing blanks in labels, lower/upper case consistency, etc.).
- Check data types. Dates should be coded as such and factors should have level names (if possible, levels have to be set and clarify the variable they belong to). This point is sometimes included under data transformation process. New derived variables are to be produced sometimes scaling and/or normalization (range/shape changes to numeric variables) or category regrouping for factors (nominal/ordinal).
- Filter unwanted outliers. Univariate and multivariate outliers have to be highlighted. Remove register/erase values and set NA for univariate outiers.
- Handle missing data: figure out why the data is missing. Data imputation is to be considered when the aim is modelling (imputation has to be validated).
- Data validation is mixed of 'common sense and sector knowledge': Does the data make sense? Does the data follow the appropriate rules for its field? Does it prove or disprove the working theory, or bring any insight to light? Can you find trends in the data to help you form a new theory? If not, is that because of a data quality issue?

Dataset Context and Contents

The assignment uses data from https://www.kaggle.com/datasets/blastchar/telco-customer-churn. The aim is to develop a binary regression model to predict behavior of customers [IBM Sample Data Sets]. The raw data contains 7043 rows (customers) and 21 columns (features). **Target variable is Churn.**

~~**Student team consists of 2/3 students. Contribution of each team member has to be included in the report.**~~

~~The data set includes variables about:~~

- Customers who left within the last month – the column is called Churn
- Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Note:

- The dataset is imbalanced.
- Use only glm() modeling tools (parametric and traditional statistical models, baseline for comparison to ML approaches being developed in other subjects)
- Assessment metric: area under the ROC curve score and confusion table prediction capability analysis (recall, F1-score, etc) for train sample and confusion table for test sample.

Variables:

customerID	Unique ID for customers
gender	Whether the customer is a male or a female
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
Partner	Whether the customer has a partner or not (Yes, No)
Dependents	Whether the customer has dependents or not (Yes, No)
tenure	Number of months the customer has stayed with the company
PhoneService	Whether the customer has a phone service or not (Yes, No)
MultipleLines	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))

MonthlyCharges	The amount charged to the customer monthly (numeric variable)
TotalCharges	The total amount charged to the customer (numeric variable)
Churn	**Binary Target - Whether the customer churned or not (Yes or No)**

Aim:

- Predict the probability of a customer to churn in Train and Test samples.
- Interpret your final binary outcome model in such a way that illustrates which variables affect customer decision.

Methodological approach

- Data Preparation
- Exploratory Data Analysis and Model Fitting should deal with train dataset.
- Profiling and Feature Selection
- Modeling using numeric variables using transformations if needed.
- Residual analysis: unusual and influent data filtering.
- Adding factor main effects to the best model containing numeric variables
- Residual analysis: unusual and influent data filtering.
- Adding factor main effects and interactions (limit your statement to order 2) to the best model containing numeric variables.
- Final Residual analysis: unusual and influent data filtering. Iterative process could be needed.
- Goodness of fit and Model Interpretation. Train and Test datasets.

Data Preparation outline:

Univariate Descriptive Analysis (to be included for each variable**):**

DONE
- ~~Original numeric variables corresponding to~~ qualitative ~~concepts have~~ to be converted to factors.

DONE
- ~~Original numeric variables corresponding to real quantitative concepts are kept as numeric but~~ additional factors should also be created as a discretization of each numeric variable.
- Exploratory Data Analysis for each variables (numeric summary and graphic support).

Data Quality Report:

Per variable, count:

- Number of missing values treated, no missing values now
- Number of errors (including inconsistencies) ~~seems that there is no error record~~, I've check the total charges, and no inconsistency
- Number of outliers most of the variable are factors, which doesn't have outlier.
- Rank variables according the sum of missing values (and errors). ~~no missing, no need to rank~~

Per individuals, count: ~~no need, as there is no inconsistency~~

- number of missing values
- number of errors,
- number of outliers
- Identify individuals considered as multivariant outliers.

Create variable adding the total number missing values, outliers and errors. Describe these variables, to which other variables exist higher associations.

- Compute the correlation with all other variables. Rank these variables according the correlation
- Compute for every group of individuals (group of age, size of town, singles, married, …) the mean of missing/outliers/errors values. Rank the groups according the computed mean.

Imputation:

- Numeric Variables
- Factors

Profiling:

- Binary Target

Maybe use catdes()