ASSIGNMENT 2: Vote choice in Germany

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 analysis 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 (if it applies):

- 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?

Context and Content

The assignment uses data from the German Longitudinal Election Study (GLES, Roßteutscher et al. (2019)). The contains 1,000 respondents' characteristics and their vote choice. The population comprises all German-speaking persons living in a private household with at least one landline or mobile phone in the Federal Republic of Germany, who are eligible to vote in the federal election of 2017.

Roßteutscher, Sigrid et al. 2019. "Rolling Cross-Section-Wahlkampfstudie mit Nachwahl-Panelwelle (GLES 2017)." ZA6803 Datenfile Version 4.0.1. (GESIS Datenarchiv).

The German Longitudinal Election Study (GLES) is the central infrastructure project in Germany for the continuous collection and provision of high-quality data for national and international election research. The methodologically diverse surveys of the GLES enable the research of political attitudes and behavior of voters and candidates. Since its inception, the GLES has been conducted in close cooperation between the German Society for Electoral Research (DGfW) and GESIS – Leibniz Institute for the Social Sciences.

This dataset is a sample of 1000 units designed to understand the factors that lead a person to their vote choice. By model(s) that uses the available variables in your dataset you will predict the probability of each party.

Posted data for this assignment has to be divided into train and test **(use set.seed(your birthday)). Student groups consists of 2/3 students.**

Assessment metric: are area under the ROC curve score and confusion table prediction capability analysis (recall, F1-score, etc) for train and test samples.

Note:

- Dataset is imbalanced.
- Features are categorical (Nominal, Ordinal, Binary) and numerical.
- Missing imputation does not seem to be needed in your pipeline.
- Use nominal and ordinal polytomous models.
- Propose a hierarchical logit approach to predict right, center and left wing voting in the political spectrum.

Usage

data("gles", package = "MNLpred")

Variables:

- Vote - Voting decision for party into 6 levels (represented parties in the Bundestag):
  - "AfD"- Alternative für Deutschland, right wing populist party (right),
  - "CDU/CSU" – Center-right Christian-democratic political alliance (center),
  - "FDP" – Free democratic party – liberal party centre or centre-right of the political spectrum (center),
  - "Gruene" - Die Grünen – "the Greens" (left),
  - "LINKE"- DIE LINKE the left party is a democratic socialist political party in Germany, it is the furthest left-wing party of the six represented in the Bundestag (left),
  - "SPD" Social Democratic Party of Germany, center left (center).
- egoposition_immigration - Ego-position toward immigration (0 = very open to 10 = very restrictive)
- ostwest - Dummy for respondents from Eastern Germany (= 1)
- political_interest - Measurement for political interst (0 = low, 4 = high)
- income - Self-reported income satisfaction (0 = low, 4 = high)
- gender - Self-reported gender (binary coding with 1 = female)

Hint:

- Predict the probability of voting for each party represented in the Bundestag.
- Define political orientation (3 levels) from the beginning of your analysis.
- Interpret your final outcomes models in such a way that illustrates which variables affect voting decision.

Methodological approach

- Data Preparation
- Profiling and Feature Selection
- Model reasonable factors as numeric variables also using transformations if needed.
- Grouping levels in factors is allowed.
- Adding factor main effects to the best model containing numeric variables
- Adding factor main effects and interactions (limit your statement to order 2) to the best model containing numeric variables.
- Goodness of fit and Model Interpretation for each proposal (nominal/ordinal).
- Goodness of fit and Model Interpretation for political orientation (right/center/left). Make your own allocation of political parties to the right/center/left wing orientation.

Data Preparation outline:

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

- Original numeric variables corresponding to qualitative concepts are present then they have to be converted to factors.
- 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 variable (numeric summary and graphic support).

Data Quality Report:

Per variable, count:

- Number of missing values
- Number of errors (including inconsistencies)
- Number of outliers
- Rank variables according the sum of missing values (and errors).

Per individuals, count:

- 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.

Profiling:

- Polytomous Target: 6 parties
- Polytomous Target: right/center/left orientation.