ASSIGNMENT 2: AIRLINE SATISFACTION

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 outliers.
- 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/teejmahal20/airline-passenger-satisfaction. The aim is to develop a binary regression model to predict behavior of customers. The raw data contains 5000 rows (customers) and 25 columns (features). **Target variable is satisfaction.**

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:

- Whether customers are satisfied with their flight– the column is called Satisfaction.
- Score columns where each customer rated specific features of the flight out of 5
- Customer account information – Travel type, Flight distance, class, customer type, delays
- Demographic info about customers – gender, age

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:

Gender	Gender of the passengers (Female, Male)
Customer Type	The customer type (Loyal customer, disloyal customer)
Age	The actual age of the passengers
Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance	The flight distance of this journey
Inflight wifi service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient
Ease of Online booking	Satisfaction level of online booking
Gate location	Satisfaction level of Gate location
Food and drink	Satisfaction level of Food and drink
Online boarding	Satisfaction level of online boarding
Seat comfort	Satisfaction level of Seat comfort
Inflight entertainment	Satisfaction level of inflight entertainment
On-board service	Satisfaction level of On-board service
Leg room service	Satisfaction level of Leg room service
Baggage handling	Satisfaction level of baggage handling
Check-in service	Satisfaction level of Check-in service
Inflight service	Satisfaction level of inflight service
Cleanliness	Satisfaction level of Cleanliness
Departure Delay in Minutes	Minutes delayed when departure

Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	**TARGET - Airline satisfaction level (Satisfaction, neutral or dissatisfaction)**

Aim:

- Predict the probability of a customer satisfaction 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**):**

- Original numeric variables corresponding to qualitative concepts 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.

Imputation:

- Numeric Variables
- Factors

Profiling:

- Binary Target