ASSIGNMENT 2: HR Analytics- Job Change of Data Scientists

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:

- 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

A company which is active in Big Data and Data Science wants to hire data scientists among people who successfully pass some courses which conduct by the company. Many people signup for their training. Company wants to know which of these candidates are really wants to work for the company after training or looking for a new employment because it helps to reduce the cost and time as well as the quality of training or planning the courses and categorization of candidates. Information related to demographics, education, experience are in hands from candidates signup and enrollment.

This dataset designed to understand the factors that lead a person to leave current job for HR researches too. By model(s) that uses the current credentials, demographics, experience data you will predict the probability of a candidate to look for a new job or will work for the company, as well as interpreting affected factors on employee decision.

Posted data for this assignment has to be divided into train and test (**https://www.kaggle.com/arashnic/hr-analytics-job-change-of-data-scientists) after selecting your own random sample of 5000 observations (use set.seed(your birthday)). Student groups consists of 2 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:

- The dataset is imbalanced.
- Most features are categorical (Nominal, Ordinal, Binary), some with high cardinality.
- Missing imputation has to be a part of your pipeline.
- Use only glm() modeling tools (parametric and traditional statistical models, baseline for comparison to ML approaches being developed in other subjects)

Variables:

- enrollee_id : Unique ID for candidate
- city: City code
- city_ development _index : Developement index of the city (scaled)
- gender: Gender of candidate
- relevent_experience: Relevant experience of candidate
- enrolled_university: Type of University course enrolled if any
- education_level: Education level of candidate
- major_discipline :Education major discipline of candidate
- experience: Candidate total experience in years
- company_size: No of employees in current employer's company
- company_type : Type of current employer
- last*new*job: Difference in years between previous job and current job
- training_hours: training hours completed
- target: 0 – Not looking for job change, 1 – Looking for a job change

Hint:

- Predict the probability of a candidate will work for the company
- Interpret your final binary outcome model in such a way that illustrates which variables affect candidate decision.

Methodological approach

- Data Preparation
- 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.

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 variables (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