Data Mining and Text Mining – Course Project

Context

On February 2017 one new potential client in Financial Services required Bip's support in the assessment on the risk of default for his credit card users. The client has already collected and structured the data of about 27.000 of its credit card subscribers, reporting some of their personal information and their aggregate payments history for the last 6 months (July to December 2016). Data has been collected in the train.csv file. Their request is to support them, by using this data, in discriminating credible subs from non-credible so to enable proactive risk management actions.

Data

The database is structured as follows:

CUST_COD	Customer code
LIMIT_BAL	Amount of credit given to the client (€)
SEX	Gender
EDUCATION	Education level
MARRIAGE	Marital status
BIRTH_DATE	Date of birth
PAY_DEC, PAY_NOV, …	History of payments for the past 6 months (Jul-Dec); The measurement scale for the repayment status is: <=0 : pay duly 1 : payment delay for one month 2 : payment delay for two months … 8 : payment delay for eight months 9 : payment delay for nine months and above
BILL_AMT_DEC, BILL_AMT_NOV, …	Amount of bill statement per month (€) for the past 6 months (Jul-Dec)
PAY_AMT_DEC, PAY_AMT_NOV, …	Amount of payment per month (€) for the past 6 months (Jul-Dec)
DEFAULT PAYMENT JAN	Default payment on January (1: Yes; 0: No)

Objectives

You have been chosen to be part of the Data Science team engaged to fulfill this task. You are asked to:

- Provide a data mining pipeline to infer on the risk of default for credit card users. Note that **the client is not interested in understanding the causes driving credit card default events, rather it is more interested in making accurate predictions**. Project compensation for Bip will be derived in form of a success-fee based on the **F_1 score** KPI calculated on the classification of an out-of-sample test dataset.
- The whole analysis pipeline must be implemented either as a Python notebook, an R notebook, or a KNIME workflow that can take the train and test data and produce the vector of predictions for the test data
- Justify your choice for the model by providing a one page (two sheets) description of your work
- Prepare a set of slides (5 maximum) to present your end-to-end analysis to the client's business department.

Outputs

You are asked to provide two outputs for this activity:

- Compile the test.csv file by providing a prediction on the default risk for each customer listed in the dataset. The classification must be a 0/1 label in the "DEFAULT PAYMENT JAN" column for each record in the dataset.
- The R/Python notebook or the KNIME you developed to generate the classification of the test.csv file
- A one page (two sheets) description of your work
- The .ppt file with the presentation of your work as described above.