Assignment Content

Top of Form

**INTRODUCTION**

A bank in your region wants to build a model to predict credit card defaulters more accurately in order to minimize money loss. For this task, we have a set of data on default payments and demographic data to help us do our task. Data is comprised in the following CSV files, each containing a set of information related to 20,000 customers.

**DATASETS**

**train\_customers.csv -**Demographic data

* + ID: ID of each client
  + LIMIT\_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
  + SEX: Gender (1=male, 2=female)
  + EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
  + MARRIAGE: Marital status (1=married, 2=single, 3=others)
  + AGE: Age in years

**train\_series.csv -**Information about payments, bill statements, and repayment status for the last 6 months.

* + ID: ID of each client
  + MONTH: The month to which data is referring (JANUARY to JUNE)
  + PAY: Repayment status in the corresponding month (-2=no need to pay, zero balance, -1=pay duly, 0=revolving credit (meaning client paid more thanthe minimum payment, but less than the total balance), 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: Amount of bill statement in the corresponding month (NT dollar)
  + PAY\_AMT: Amount of previous payment in the corresponding month (NT dollar)

**train\_target.csv -**If the customer default in the next month

* + DEFAULT\_JULY: Default payment in July (1=yes, 0=no)

**test\_data.csv** - For Part 3

**submission\_features.csv** - For Part 3

**train\_data.csv** - Backup for Part 2 and 3.

**SUBMISSION DETAILS**

At the end of the assignment, you must upload the following files:

* + ***<surname><name>\_Assignment1.ipynb -***The notebook for this assignment. This notebook contains code for all parts. A template is provided to the student.
  + ***<surname><name>\_A\_train.csv -***This is the resulting CSV after applying all transformations exposed in questions A.1, A.2 and A.3.
  + ***<surname><name>\_C1\_submission.csv* -**This is the answer to question C.1. More details are in the specific section.
  + Questions C.2 and C.3 must be answered on Blackboard. If no answer is found there, the question will be evaluated as 0

**GENERAL RULES**

Please, bear in mind the following rules

* + All submitted files must be in the specified format or they will be ignored (and not evaluated)
  + Some questions (C.2 and C.3) must be answered on Blackboard. If no answer is found, the question will be evaluated as 0.
  + Independently of the submission model, all questions must be answered in the notebook of the assignment
  + After the deadline, no additional submissions can be made. If there are partial submissions of missing files, those will cannot be provided after the deadline.
  + **Avoid cheating**. This is an individual assignment. If any act of cheating (partial or total) is detected between two assignments, both will be evaluated as 0.

Bottom of Form

Top of Form

**submission\_features.csv**

**test\_data.csv**

**train\_customers.csv**

**train\_data.csv**

**train\_series.csv**

**train\_target.csv**

Bottom of Form

Top of Form

**IndividualAssignment.ipynb**

Bottom of Form

Top of Form

**PART A: DATA ENGINEERING (3 points)**

In this part, we're going to prepare data for further analysis. The goal is to transform the separated files into one analytic table with one row per customer (ID).

**SUBMISSION DETAILS**

Upload a CSV called ***<surname><name>\_A\_train.csv***with all transformations applied.

Bottom of Form

Top of Form

Question 1

**2 Points**

**Load the*train\_series.csv* file and pivot the time-dependent columns (PAY, BILL\_AMT, and PAY\_AMT) to build one column of these per month. *(Ex. PAY\_JUNE, PAY\_MAY, ..., BILL\_AMT\_JUNE, ...)***

Bottom of Form

Top of Form

Question 2

**0.5 Points**

**Load *train\_customers.csv* and include the pivoted table of time series data. The resulting table must be a table with one row per customer (ID)**

Bottom of Form

Top of Form

Question 3

**0.5 Points**

**Include information from the target (*train\_target.csv*) on the table**

Bottom of Form

Top of Form

**PART B: EXPLORATORY DATA ANALYSIS (2 points)**

Let's do some exploratory data analysis over the training data obtained from the previous part.

**SUBMISSION DETAILS**

No additional files should be uploaded. All code, graphics, and answers must be in the main notebook of the assignment.

**NOTE:**Exceptionally, if you were not able to complete Part 1, you can use the ***train\_data.csv*** dataset for this Part.

Bottom of Form

Top of Form

Question 4

**0.5 Points**

**BAR PLOT: Plot the distribution of defaulters given the Education. Comment the results.**

Bottom of Form

Top of Form

Question 5

**0.5 Points**

**BOX PLOT - Plot the distribution of `LIMIT\_BAL` per defaulters and non-defaulters. Comment the results.**

Bottom of Form

Top of Form

Question 6

**1 Point**

**Include 3 more interesting insights extracted from the data. Provide an explanation of the choices made and the results obtained.**

Bottom of Form

Top of Form

**PART C: DATA ANALYTICS (5 points)**

In this part, we're going to build a machine learning model to estimate the probability of default of the customers. For this part, you will need two additional datasets:

* + ***test\_data.csv*** - Contains new data from 9000 customers with the same structure as the training data from Part 1. It contains all the features and the target already joined. You can use this data set for testing purposes (or you can build your own test set from the training set). Anyway, keep it because you will need it for question C.2
  + ***submission\_features.csv*** - This contains data associated with unseen customers, but without the target.

**SOME CONSIDERATIONS**

* + Try several models to select the best suited for this problem
  + Don't forget about hyperparameter tuning to select the best configuration for the model
  + Don't forget to check the different metrics in each CV round to build a robust model

**NOTE:**Exceptionally, if you were not able to complete Part 1, you can use the ***train\_data.csv*** dataset for this Part.

Bottom of Form

Top of Form

Question 7

**1.5 Points**

**Build a machine learning model to estimate the probability of default.**

**SUBMISSION DETAILS**

Once you have your best model, make new (soft) predictions over the *submission\_features.csv*. Store your predictions in a CSV (**<surname><name>\_C1\_submission.csv**) with two columns:*ID* and *probability*. The *ID* must contain the customer ID and the *probability* must contain your model's probability output for class 1 (default). For example:

ID,probability

23,0.87

1254,0.45

2654,0.78

...,...

345,0.12

**GRADING CRITERIA**

The following criteria will be considered

* + With the submitted results, the **Area Under the Curve of the ROC (AUC-ROC)** will be calculated using the submission set
  + The grading criteria will consist of a linear function from the worst ROC-AUC to the best ROC-AUC
  + **100%** - Your ROC-AUC score is the best
  + **50%** - Your ROC-AUC score is the worst
  + If your ROC-AUC is lower than the professor's dumb model, the grade will be **25%**.
  + If the submission is not in the correct format, or the professor can't evaluate the exercise the grade will be **0%**.

Bottom of Form

Top of Form

Question 8

**1 Point**

**Which five of 1000 clients from the submission set are most likely to repay the loan if it were granted?**

**SUBMISSION DETAILS**

Paste the IDs in below

Bottom of Form

Top of Form

Question 9

**2.5 Points**

Now the bank wants to optimize the decision-making process by establishing the optimal threshold for the model in order to effectively take the decision about when to issue the credit and when not. So, taking into account the following numbers:

* + A customer who received a loan but doesn't repay **costs 5000$** to the bank
  + A customer who receives a loan and repays, make a **profit** of **1000$** to the bank
  + If the credit is not issued, then there is no profit or loss

**What's the best threshold you should consider to maximize the profit?**

*NOTE: You can use your test set to make your calculations.*

**SUBMISSION DETAILS**

Paste the optimized threshold below

**GRADING CRITERIA**

The following criteria will be considered

* + With the provided threshold, the *Profit* will be calculated over the submitted results in *C.1*
  + The grading criteria will consist of a linear function from the worst Profit to the best Profit
  + **100%** - Your Profit is the best
  + **50%** - Your Profit is the worst
  + If you don't make any profit, but losses, the grade will be **25%**. (And probably you will be fired of the company...)
  + If the submission is not in the correct format, or the professor can't evaluate the exercise the grade will be **0%**.

Bottom of Form