### Week 12: Final Team Presentations (TP02)

**Time: 15-min**

1. An overview of the competition - what's the objective? where does the dataset come from? what are the key features?

**Introduction**:

Modern life counts on the convenience of a credit card to make daily purchases. How do card issuers know we’ll pay back what we charge?

Credit default prediction is central to managing risk in a consumer lending business. Credit default prediction allows lenders to optimize lending decisions, which leads to a better customer experience and sound business economics.

**Objective**:

The objective of this competition is to predict the probability that a customer does not pay back their credit card balance amount in the future based on their monthly customer profile.

Dataset details:

The target binary variable is calculated by observing 18 months performance window after the latest credit card statement, and if the customer does not pay due amount in 120 days after their latest statement date it is considered a default event.

Files

train\_data.csv - training data with multiple statement dates per customer\_ID

train\_labels.csv - target label for each customer\_ID

test\_data.csv - corresponding test data; your objective is to predict the target label for each customer\_ID

sample\_submission.csv - a sample submission file in the correct format

**Aggregated features:**

The dataset contains aggregated profile features for each customer at each statement date. Features are anonymized and normalized, and fall into the following general categories:

* D\_\* = Delinquency variables
* S\_\* = Spend variables
* P\_\* = Payment variables
* B\_\* = Balance variables
* R\_\* = Risk variables

with the following features being categorical:

['B\_30', 'B\_38', 'D\_114', 'D\_116', 'D\_117', 'D\_120', 'D\_126', 'D\_63', 'D\_64', 'D\_66', 'D\_68']

Your task is to predict, for each customer\_ID, the probability of a future payment default (target = 1).

Note that the negative class has been subsampled for this dataset at 5%, and thus receives a 20x weighting in the scoring metric.

**2. A brief critique of select Notebooks in Python or R on this competition available in the public domain - The team should critically evaluate other people's published work using concepts learned from the Machine Learning 1 and Machine Learning 2 coursework.**

Some of the teams used AI models in python which is not the goal of this assignment.

80.199Team visa used LightGBM is a gradient boosting framework based on decision trees to increases the efficiency of the model and reduces memory usage. And catgboost: CatBoost is an open-source software library developed by Yandex. It provides a gradient boosting framework which among other features attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm. GBDT models (Gradient-boosted decision trees are a machine learning technique for optimizing the predictive value of a model through successive steps in the learning process) . I liked the presentation of the transformation of the data and model.

80.19We used logistic regression to ensemble ~60 sets of predictions with diversity obtained from different models, feature sets, and data views. Models included the usual mix of gradient boosters (LGBM, LGBM dart, XGBoost, CatBoost), MLPs, TabNet, and Transformers with a few special tricks. Data views included aggregated data, fully flattened data (13 raw features per customer), sequential data (Transformer), and augmented data (shifted statements forward by 1 and appended to original). Very detailed feature information to follow -- they are split across our 3 different team members' different models, but this aims to be a complete compilation of what we all used.

Diagram

Description automatically generated

**3. Based on what the team has researched, the team should create their own solution in R to the competition and explain why they believe the solution is the best, and what is new or different from published Notebooks in some conceptually rigorous way based on the Machine Learning 1 and Machine Learning 2 coursework. You must code your solution in R.**

The team should create a Github page to host their work, including code and/or presentation (include PPT only if needed). The captain will submit the link to the TP2 GitHub page in Blackboard for grading by the time of the scheduled presentation.  
  
**Grading rubric (details available in Blackboard)**  
**Problem Statement (10%)**: Clear, concise, accurate and focused statement of the problem being solved  
**Critiques (20%)**: Clear, accurate and thoughtful critiques of existing body of work that focus on key issues of algorithms, datasets, or code, and/or their reproducibility issues.  
**Solution (50%)**: Thoughtful and comprehensive solution that addresses the problem statement in a logical manner and issues identified in the critiques. Proper data processing and exploration and clear rationale for algorithm choices.  
**Reproducibility (10%)**: Comprehensive, clear, and focused explanation of the data exploration and processing work, rationale for the algorithm choices, and documentation of your code, based on this [Machine Learning Reproducibility Checklist](https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf). Be sure to document rationale for WHY you perform certain tasks/made certain decisions, and not just WHAT you did.  
**Presentation (10%)**: Clear, well rehearsed, and concise presentation with laser focus on the key topics, key issues and key elements of the solution. Free of errors.  
  
Each team should select a unique competition (i.e., no two teams, across the two class sections, can select the same competition). There are a total of 21 teams in Spring 2022. Topic choices are first come first serve. As soon as the team decides on a competition, the team captain should email the instructor to claim the competition. The instructor will post team topic claims on this webpage here so everyone can see what's available, what's taken, etc. The following is a list of potential topics. Teams are welcome to select a competition not on the list, as long as the topic is approved by the instructor.