Machine Learning in Finance
Group Project

- This is a group project due on **08/01/2026**
- Groups should have **3-5** members each.
- All projects should be completed using **Jupyter notebooks** which will be delivered at the end of the project. Each task should have an independent jupyter notebook, to assist in the grading.
- Please read the instructions as you also need to deliver a **video**.

Project Context

Home Credit is a financial services company that provides loans to customers often underserved by traditional banking institutions. The goal is to use machine learning to assess the risk of default at the loan application stage, enabling better decision-making and optimized loan offerings. Understanding which customers are likely to default is crucial for reducing financial losses and improving operational efficiency.

Project data

You have access to the Home Credit Default Risk dataset, along with a data dictionary to help you understand the column titles. However, you will need to explore the data yourself—not all columns will be useful, and some may require significant preprocessing. Your tasks will include:

- Cleaning the data (e.g., handling null values, resolving data type issues, addressing outliers).
- Dealing with categorical variables and feature engineering.
- Exploring which features are useful for predictive tasks and avoiding features that may leak data.

Project Tasks

Task 1 Data Cleaning and Exploration

A. Explore and understand the data. Use the data provided only.
B. Convert the data into features. Think about types of variables and the requirements of the algorithms you will use for prediction.
C. There are two tasks below, they may require different data. Decide or create the features you will use for each task. All features might not be useful or relevant or may leak data. For each feature think, would this be available at the time the machine learning model needs to be

applied. **Warning** if your metrics like recall on charge off are very high, you may be leaking data.

Please deliver a jupyter notebook that shows your working for task 1. This workbook should read in the initial dataset and write datafiles that will be read in by the notebooks for the remaining tasks.

Task 2 Classification Model to Predict Payment Difficulties

By predicting which loans are likely to have payment difficulties at the time of application, Home Credit can take **preventive action and only issue loans to customers that will pay off their debts**, thereby reducing losses and improving their service to both borrowers and investors.

A. Build the overall "best" classification model to predict if a loan will be charged off.
   - You will need to decide how you define and demonstrate which model is "best".
   - You can use algorithms outside of those seen in class.
   - Comment on any limitations this model might have when used by management for its desired purpose.

   Important you will mostly be graded on your ability to apply the machine learning approaches we have learned in class properly. So please include the work that you test and not just the final best model.

B. Assume that Lending Club wants to make sure that they identify all loans that will have payment issues within reason.
   - How can you increase the ability of the classifier to do this? What is the best classifier you can find to achieve this goal.

Please deliver a jupyter notebook that shows your working for task 2. This workbook should read in the data files generated in task 1.

Task 3 Customer Segmentation

Understanding the diverse needs and behaviours of customers is crucial for targeted marketing, risk assessment, and product development. Customer segmentation can lead to more personalized loan products and better customer satisfaction.

A. Apply different approaches to unsupervised learning to determine different customer segments.

    - You will need to decide which features to use, some of them may be repetitive or very similar for this specific task. Think how this would effect the segmentation.
    - You will need to decide how many segments to split the data into and demonstrate how you made this decision.
    - Comment on any limitations this model might have when used by management for its desired purpose.

    Important you will mostly be graded on your ability to apply and interpret the machine learning approaches we have learned in class properly. So please include the work that you test and not just the final best model.

B. Decide the best segmentation, justify why this is the best segmentation.
C. See if you can see any pattern in the different segments, if you had to explain the segments to management what would you name each one and what properties are they associated with..

Please deliver a jupyter notebook that shows your working for task 3. This workbook should read in the data files generated in task 1.

Computation

Dealing with computation constraints is a big part of machine learning, especially as we start to deal with real world data. If at any point you are trying to run an algorithm or approach that is taking a long time to solve on google colab you have a couple of options:

- You can try running it on your laptop directly, in case your laptop or computer is faster.
- You can try to reduce the complexity of the model.
- You can reduce the complexity of the data by reducing the number of columns you use at a time.
- You can just take a smaller sample of the data (less rows) and apply the algorithm / model to this.

Deliverables

- You need to **deliver all Jupyter notebooks and code used to complete the projects**.
    - You should make use of the markdown cells in the Jupyter notebooks to explain clearly what you are doing in each code cell and why.

    - You should also use these cells to provide your answers to the questions such as showing which is your best classifier and how you decide it was best.
    - Jupyter notebooks should be "solved" so I can see the results of running the code without having to run it myself.
- **You should record a video explaining your work and results**, you can go through your Jupyter notebook and explain everything you did and why you made the decisions you did, present all the problems you resolved, and your final answer to each of the questions.
    - An easy way to record this can be to just record a zoom call together.
    - Remember the idea is to talk through your working/notebooks, you don't need any slides.
    - This video should be less than 15 minutes, each group member should explain the work they completed.
    - The point of the video is to help demonstrate your thinking and understanding of the problem and the application of Machine Learning, please try to demonstrate this in the video.

Grading

- **Important** you will mostly be graded on your ability to apply the machine learning approaches we have learned in class properly. So please include the work that you test and not just the final best model.
- You might receive extra points for demonstrating the ability to apply approaches not covered in the class.
- What your final model is (and how well it performs) matters much less than your ability to correctly apply different approaches, explore the problem, and understand the final results and model. Please try to demonstrate in the notebooks and the video.
- Well organized Jupyter notebooks and code will help ensure that your intentions are communicated and are likely to increase your grade.

For any questions please contact: iscott@novaims.unl.pt