esade

RAMON LLULL UNIVERSITY

Business Case with Machine Learning

Supervised Learning: Classification Algorithms

MSc Business Analytics

Prof. Marc Torrens

TAs: Gal·la Garcia-Castany

The grade is based on 3 factors

1. Technical Solution & Analysis – 40%

- The student has appropriately followed the instructions.
- The student has applied the methodology and basic ML practices explained in class.

2. Business Interpretation & Recommendations – 40%

- The student is able to interpret ML models and their business implications.
- The student is able to propose business recommendations based on the findings.

3. Design of Deliverable – 20%

- The student communicates his/her results clearly, briefly and with a good design.

Deliverables

1. PDF

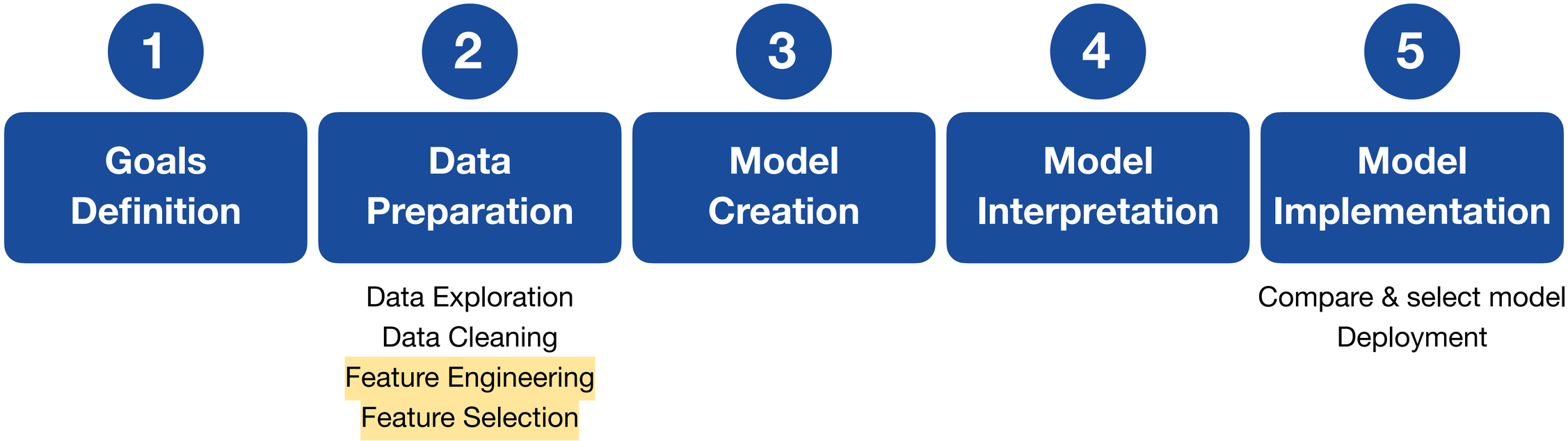
- The objective of the slide deck is to present a business presentation explaining what has been done, why it was done, and its economic impact. Answering the questions of the case.

2. Jupyter Notebook

- Full code used with clear documentation to facilitate reproducibility and further analysis, addressing all the questions in the assignment.

About the case

- The objective is to apply the ML lifecycle steps to the Lending Club case.



- Lending Club (www.lendingclub.com) is a peer-to-peer lending platform that provides loans to individuals and businesses. Understanding credit risk is crucial for financial institutions to minimize defaults and maximize profitability.

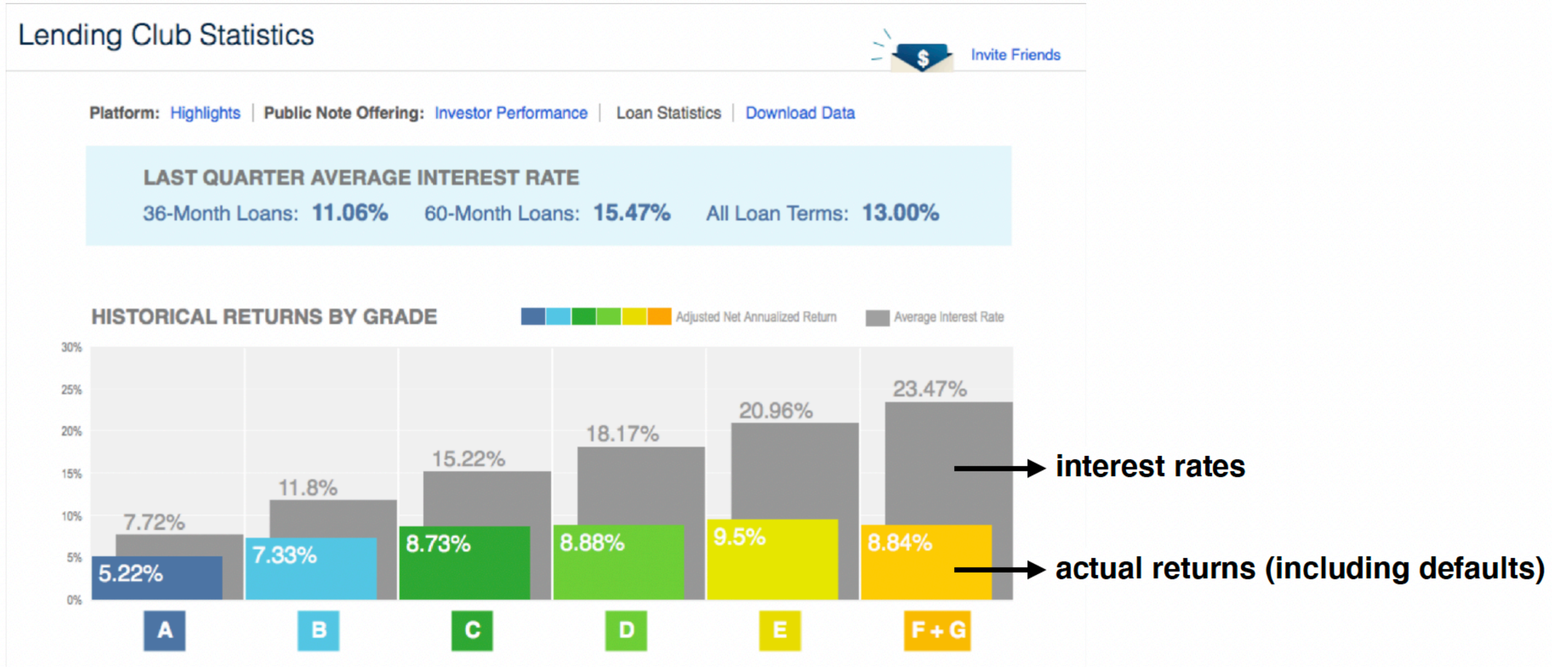
About the case

- US citizens have a credit score that goes from A to G representing the creditworthiness of a person. Lending Club assigns loan grades (A to G) based on multiple factors, including FICO scores, income, and debt-to-income ratio (DTI). Each grade corresponds to a different interest rate range, where A-grade loans have the lowest interest rates and G-grade loans have the highest.

About the case

Find the characteristics to the loan to see if the loan will be repaid or not

- Relationship between credit scores and interest rates follows a general risk-based pricing model:
  - Higher Credit Scores → Lower Interest Rates. Borrowers with higher credit scores are considered low risk (i.e., less likely to default). As a result, Lending Club (or any lender) offers them lower interest rates to attract them.
  - Lower Credit Scores → Higher Interest Rates. Borrowers with lower credit scores are considered high risk (i.e., more likely to default). To compensate for this risk, lenders charge them higher interest rates to cover potential losses.



About the case

- In a loan default prediction problem, the costs associated with errors are highly asymmetric, making the FP/FN tradeoff particularly interesting and practically relevant.
- False Positives vs. False Negatives
  - False Positives (FP):
    - These occur when the model predicts that a loan will default when, in reality, the borrower repays it. The consequence for the lending institution is a missed opportunity, as a creditworthy borrower was wrongly classified as risky and denied a loan. Over time, such errors can reduce the institution's customer base and profitability.
  - False Negatives (FN):
    - These happen when the model predicts that a borrower will repay the loan, but they actually default. This leads to direct financial losses for the institution, as it extends credit to high-risk borrowers. Accumulating such losses across multiple loans can significantly impact the institution's overall risk exposure.
- Cost sensitivity
  - In loan prediction, the cost of a FP (losing a potential profitable customer) is generally much higher than the cost of a FN (granting a loan that defaults). This asymmetry requires a careful balancing act. Institutions might prefer to err on the side of caution (minimizing FNs) even if it means a higher FP rate, or vice versa, depending on their risk appetite and business strategy.
- Threshold Tuning: the business implementation objective is to adjust the decision thresholds to optimise the tradeoff.

About the case

- Investment strategies based on credit scores and interest rates:
  - 1. Conservative Strategy – Low Risk, Stable Returns
    - Target Loans: A & B grade loans (High credit scores, Low interest rates)
      - Risk Level: Low / Expected Returns: Moderate but stable / Default Risk: Very low
  - 2. Balanced Strategy – Moderate Risk, Optimized Returns
    - Target Loans: B, C, and D grade loans (Medium credit scores, Moderate interest rates)
      - Risk Level: Moderate / Expected Returns: Higher than conservative, with manageable risk / Default Risk: Medium
  - 3. Aggressive Strategy – High Risk, Maximum Yield
    - Target Loans: D, E, F grade loans (Low credit scores, High interest rates)
      - Risk Level: High / Expected Returns: Very high, but with substantial risk / Default Risk: High

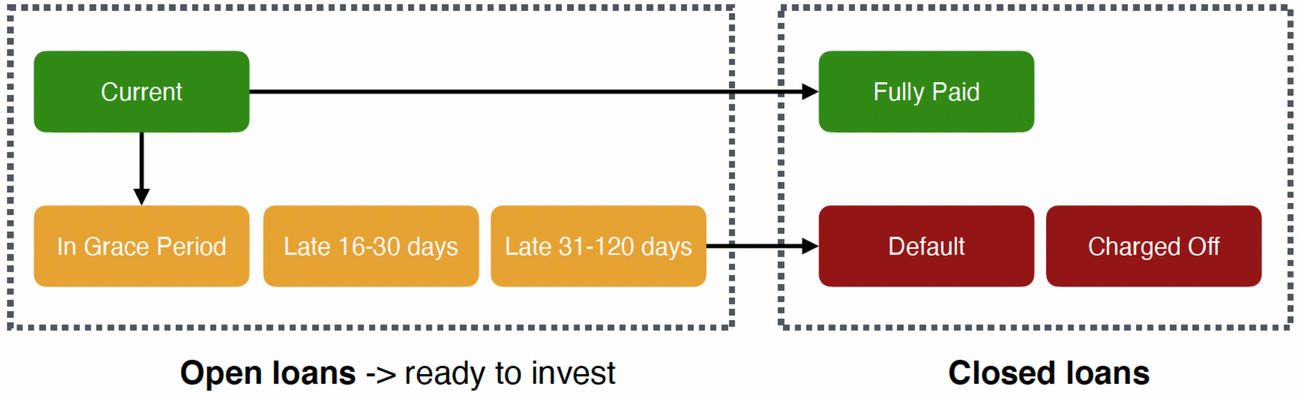
Compare different investment strategies on the different credit grades and assess their profitability based on default rates, interest returns, and risk exposure. Your goal is to determine the optimal strategy to maximize returns while effectively managing risk.

Data Preparation

- 1. Load the dataset “Loan_Lending_Club.csv” and conduct an initial exploration of the data.
- 2. Handle missing values appropriately.
- 3. Determine which categorical variables should be included in the model based on their predictive value.
- 4. Explore ways to preprocess date columns to improve predictive modeling.
- 5. Analyze correlations between numerical features and select appropriate independent variables for the model.
- 6. Perform feature engineering (feature selection, feature transformation, feature creation, and evaluation)?
- 7. Apply necessary data transformations, such as normalization or standardization.

Model Creation

- 1. Define the target variable (y), considering that loans can be either “closed” or “open” and categorized as good (green) or bad (red/orange).



- 2. Train different classification models using only the “closed” loans and compare their performance on training and testing data.
- 3. (Bonus) Explore techniques to rebalance the dataset and assess improvements in model performance.

Model Interpretation

- 1. Identify the most relevant performance metrics (e.g., precision, recall) for evaluating your model.
- 2. Generate a confusion matrices for the different models.
- 3. Interpret the models' results, focusing on key insights and implications (based on the different investment strategies).

Model Implementation

- 1. Compare the confusion matrices of different models.
- 2. Analyze the business impact of false positives (FP), false negatives (FN), true positives (TP), true negatives (TN), and true positives (TP) - matching assumptions.
- 3. Modify the decision threshold to balance cost sensitivity effectively.
- 4. Determine the business implications of the best-performing model based on its confusion matrix results.
- 5. Apply the model developed with the “closed” loans to the “open” loans, and estimate the expected business impact.
- 6. Ultimately, compare different investment strategies on the different credit grades and assess their profitability based on default rates, interest returns, and risk exposure. Determine the optimal strategy to maximize returns while effectively managing risk.

esade
RAMON LLULL UNIVERSITY

Business Case with Machine Learning

Unsupervised Learning: Clustering Algorithms

Business Analytics

Prof. Marc Torrens

TAs: Gal·la Garcia-Castany