Loan Default Prediction
Predictive Analysis of Credit Risk in Banking

Applying Machine Learning to Anticipate Borrower Behavior and Reduce Financial Risk

Prepared by: Carlos YFRAZIN and Jean Guetcheen CHARLES
Supervised by: MyAkademi
Date: October 2025

Start Slide

01

02

03

Context and Project Objective

01

02

03

The banking sector faces a major challenge: accurately assessing credit risk before granting a loan. A high default rate can threaten the financial stability of any institution. Today, thanks to Data Science, it is possible to leverage historical loan data to identify high-risk borrower profiles. This project seeks to address two central questions: how can we detect, prior to loan approval, whether a client is likely to default, and which factors are most influential in determining repayment behavior? The objective is to build a reliable predictive model to estimate the Loan_Status (0 = repaid, 1 = default), identify the key explanatory variables that drive loan outcomes, and provide actionable recommendations to help the bank strengthen its credit risk management.

Approach and Methodology – CONTENT

The study follows a fully data-driven approach, structured around five key stages. First, the data were imported and explored to understand the structure and assess the quality of the dataset. This phase allowed the identification of variable types, their distributions, and the detection of missing or inconsistent values. Next, a thorough data cleaning and preprocessing process was carried out, including the removal of duplicates, handling of missing values, and correction of data types to ensure consistency and reliability.

The third stage, known as feature engineering, involved creating new and more meaningful variables for analysis, such as the Debt_to_Income_Ratio, an indicator that reflects an applicant's level of indebtedness relative to income. Afterwards, several supervised machine learning algorithms were trained and compared — namely Logistic Regression, Random Forest, XGBoost, and LightGBM — to evaluate their ability to predict loan repayment status. Finally, the models were assessed and interpreted using key performance metrics such as accuracy, recall, F1-score, and ROC-AUC, while identifying the most influential variables affecting credit default risk.

Data Exploration CONTENT

01

02

03

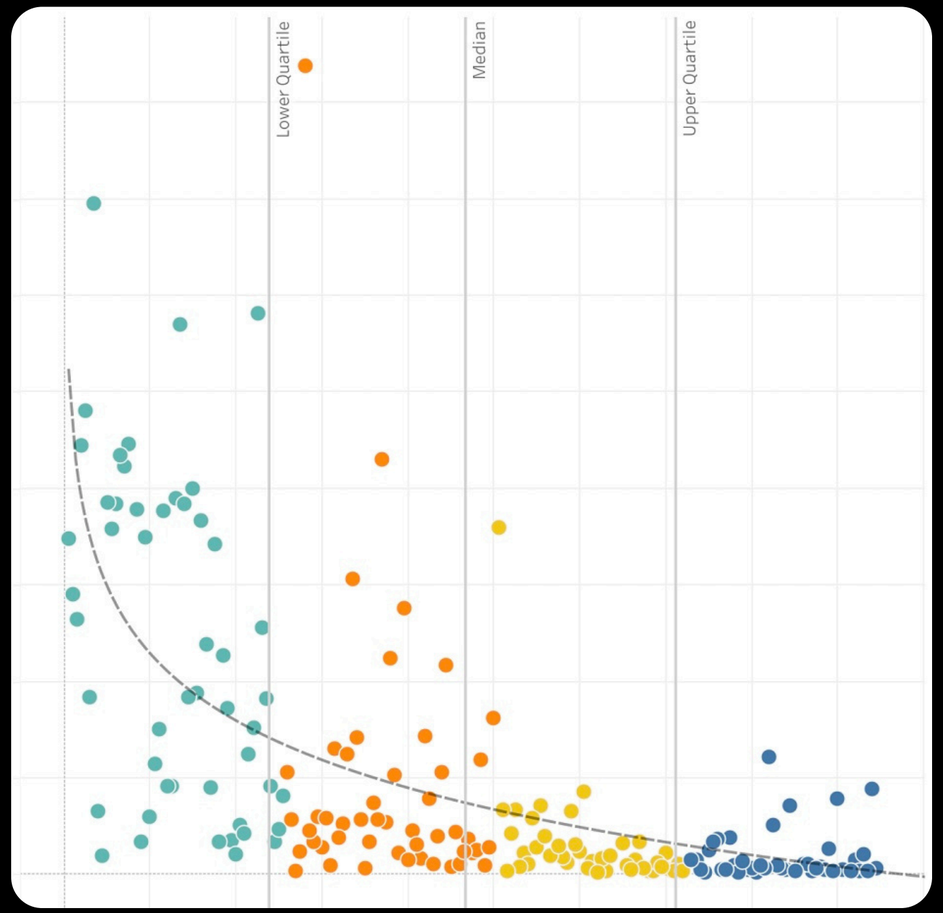
Dataset Overview:

- Size: 67,463 observations and 35 variables.
- Source: Internal database simulating historical bank loan records.
- Target variable: Loan_Status (0 = repaid, 1 = default).
- **Distribution**:
  - 91% of borrowers are good payers (class 0).
  - 9% of borrowers are defaulters (class 1).
-

Data Preparation CONTENT

Cleaning and Transformation:

- Removal of duplicate entries and systematic handling of missing values.
- Conversion of categorical (text) variables into numerical form using LabelEncoder.
- Normalization of numerical data to ensure consistency across different scales.



Training and Model Results – CONTENT

Tested Models:

1. Logistic Regression – simple baseline and interpretable.
2. Random Forest – non-linear model, robust to outliers.
3. XGBoost – high-performance boosting algorithm.
4. LightGBM – fast and efficient model for large datasets.

Analysis:

- Models without class balancing are biased towards the majority class.
- After oversampling, XGBoost and LightGBM begin to partially identify high-risk profiles.

Most Important Variables – CONTENT

Top 10 Variables Influencing Default:

1. **Recoveries**
2. **Collection Recovery Fee**
3. **Revolving Balance**
4. **Total Revolving Credit Limit**
5. **Interest Rate**
6. **Total Current Balance**
7. **Home Ownership**
8. **Total Received Interest**
9. **Debt to Income**
10. **Loan Amount**

Analysis:

→ Clients with high interest rates, significant debt, or a low repayment history exhibit a higher risk of default.

Strategic Recommendations – OTHERS

1. Optimize Lending Policy:
2. Adapt interest rates and credit terms based on the predicted risk score.
3. High-risk clients should be offered more cautious terms or additional guarantees.

1. Implement an Automated Scoring System:
2. Deploy the XGBoost model as an automatic pre-evaluation tool for loan applications.

01

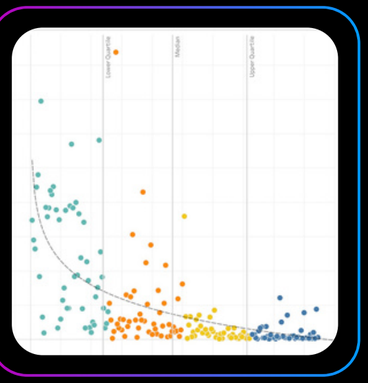
1. Rebalance Training Data:
2. Use techniques such as SMOTE to improve the model's sensitivity to the minority class (defaults).

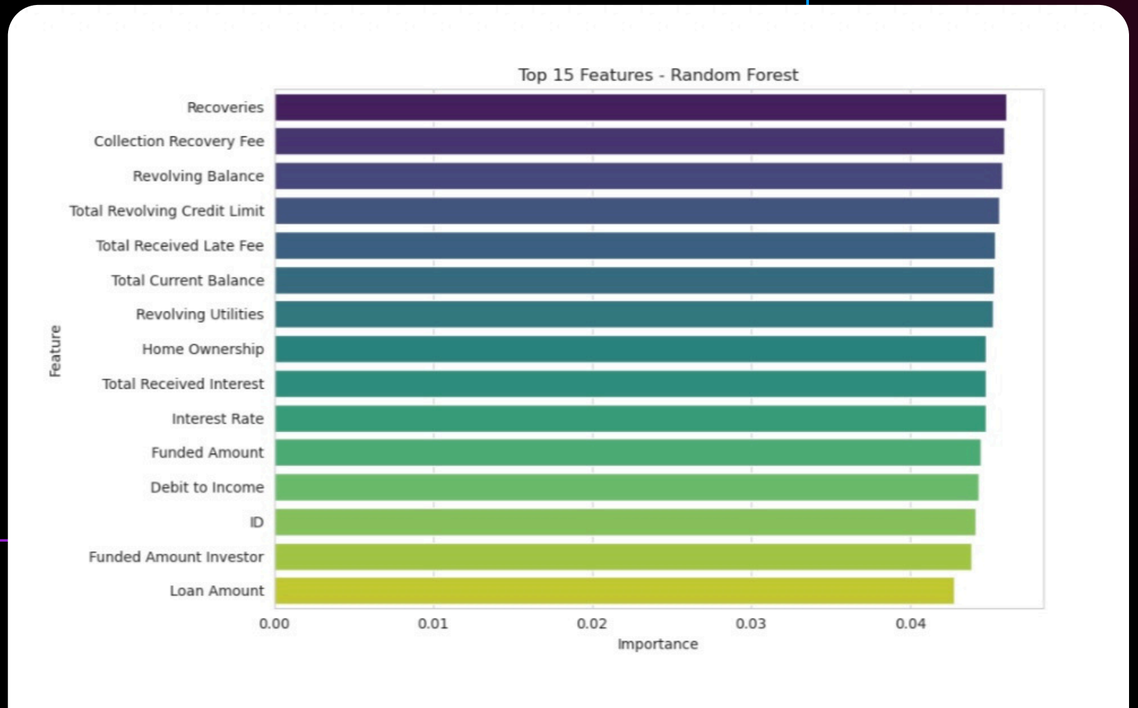
02

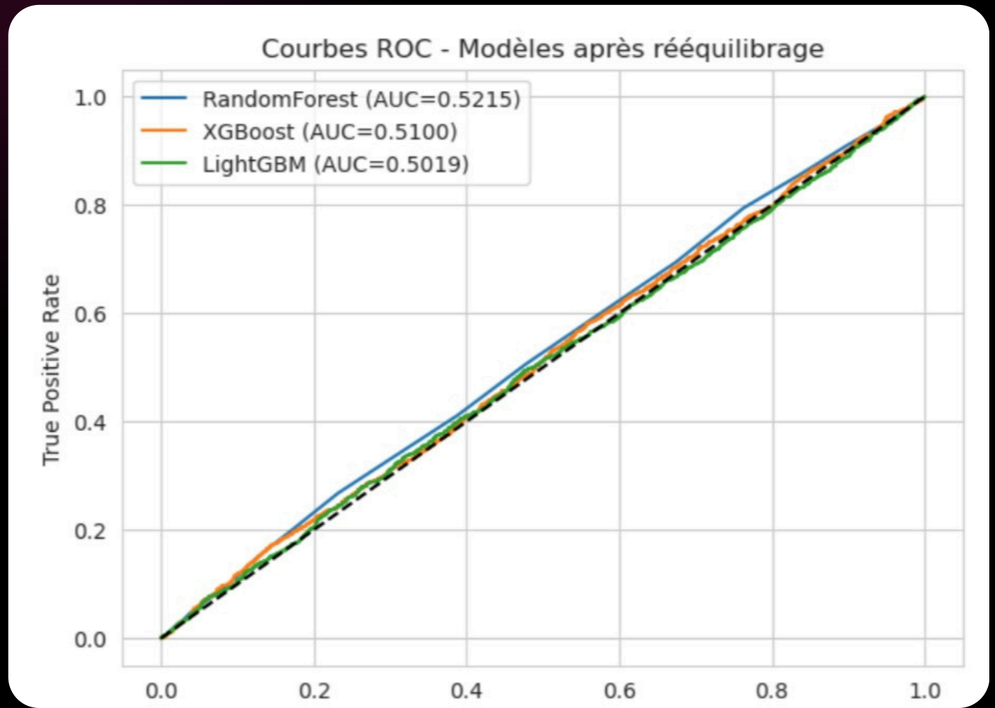
1. Proactive Monitoring:
2. Set up alerts for clients with abnormally high debt-to-income ratios.

03

Latest Portfolio







Conclusion and Acknowledgements

This project enabled us to:

- **Demonstrate the relevance of Machine Learning in credit risk assessment.**
- **Identify the most influential financial variables in loan defaults.**
- **Highlight the importance of handling class imbalance.**
- **Propose concrete recommendations to improve lending policies.**

Personal Conclusion:

This project allowed me to deepen my skills in Data Science applied to finance, while developing a business-oriented mindset focused on decision-making.

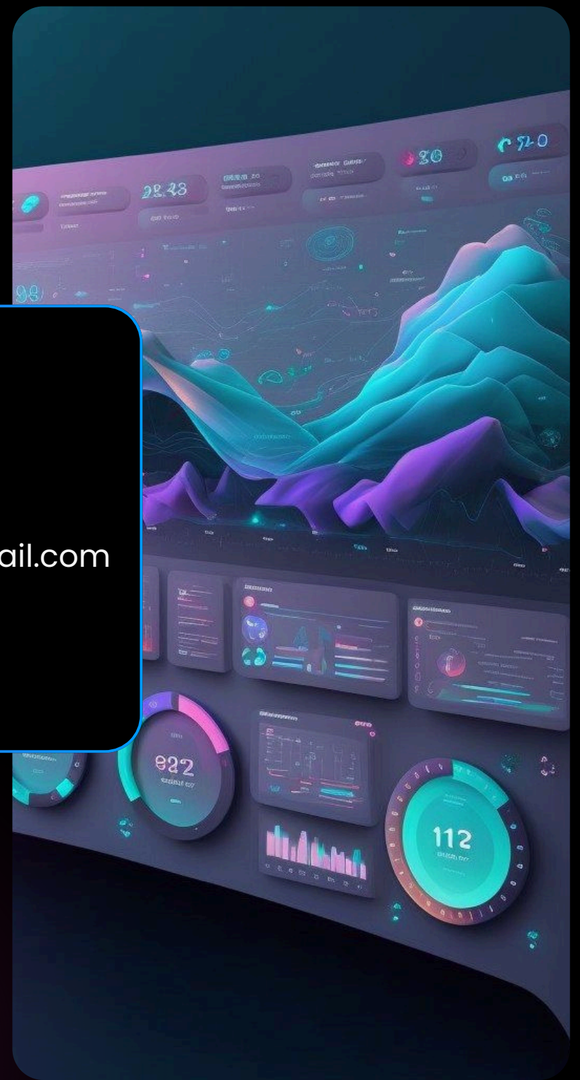
Get in Touch

Phone Number
4226-8480/42080386

Email
Carlos.yfrasin@uniq.edu
jeanguetcheencharles23@gmail.com

Social Media
Carlos yfrasin

Social Media
Jean Guetcheen Charles



01

02

03

Thank You

01

02

03

2025 Presentation