A Data-Driven Approach to Smarter Lending

Using Feature Engineering & Predictive Modelling

Name: Meghana Murali
Date: August 5th, 2025

The Challenge: Seeing the Full Picture of Risk

- **The Problem:** Credit decisions often rely on simple application data, but a customer's true risk profile is often hidden in complex data and subtle behavioural patterns.

- **Our Two-Part Goal:**
  - **Feature Engineering:** Unlock the predictive value hidden in raw, semi-structured credit report data.
  - **Predictive Modelling:** Use historical application data to build a reliable model that can accurately predict loan default.

From Raw JSON to Actionable Risk Features

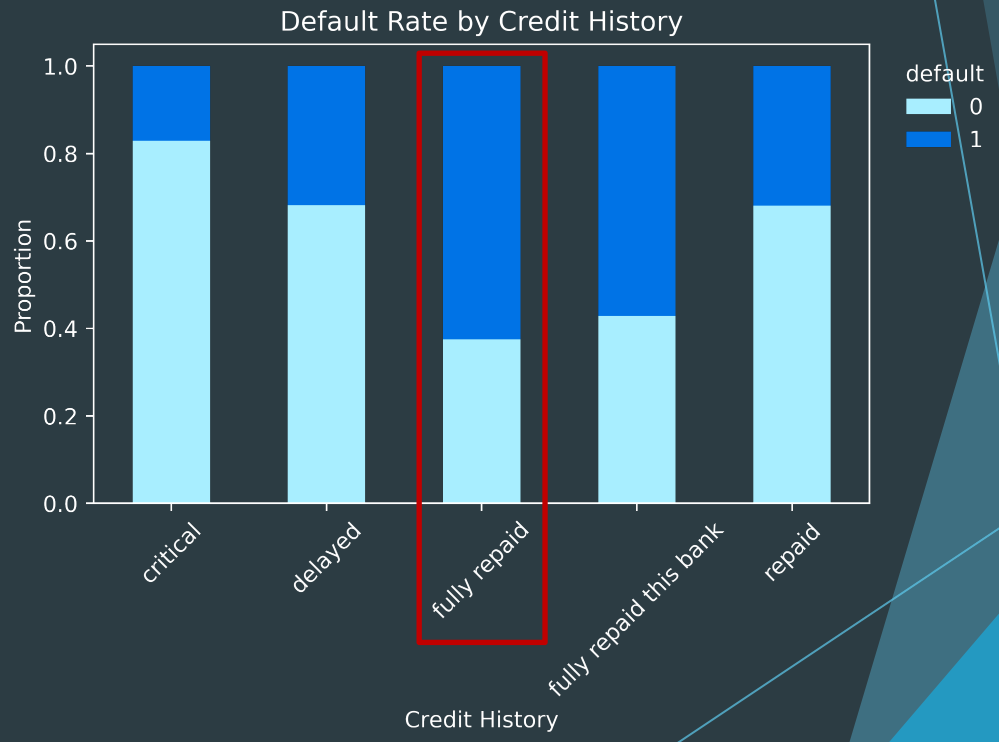
The first task was to process complex credit report data to create a rich set of predictive features.

- **Analysed Raw Data:** Deconstructed deeply nested JSON credit reports to identify key information sources like repayment history, delinquencies, and recent credit inquiries.
- **Engineered 58 New Features:** Built a reusable Python function that extracts 58 distinct features across four key categories:
  - Demographics & Stability
  - Delinquency & Negative Events
  - Credit Seeking Behaviour
  - Repayment Behaviour
- **The Outcome:** A structured, feature-rich dataset ready to enhance any credit model.

Predictive Modelling: Key Findings & Final Model

The second task was to use a historical loan dataset to train and validate a predictive model.

- **Deep Data Analysis:** Analysed the provided 'credit' file to identify the strongest predictors of default.
- **Critical Finding:** Uncovered a highly counterintuitive pattern in the credit_history data, which suggests a potential data definition issue that requires business consultation before this feature can be fully trusted.
- **Ethical Modelling:** Explicitly **excluded sensitive features like Gender** to build a fair and responsible lending model, in line with modern best practices.
- **Final Model:** Trained and tuned a powerful **XGBoost machine learning model** to achieve the best predictive performance.



How Well Does the Model Work?

Our final, tuned model is a powerful tool for differentiating between high-risk and low-risk applicants.

- **The model successfully identifies 65% of all actual defaults:**
  - This high Recall allows to proactively prevent the majority of potential credit losses before they happen.
- **It's Reliable When Flagging Risk:**
  - When the model flags an applicant as "high-risk," it is correct **53% of the time.**
- **It's a Strong Predictor Overall:**
  - With a **ROC AUC score of 0.7845**, the model demonstrates a strong and reliable ability to separate good customers from bad ones.

Sr. No	Metric	Value
1	ROC-AUC Score	0.7845
2	Precision (default = 1)	0.53
3	Recall (default = 1)	0.65

Recommendation: Implement a Risk-Based Lending Strategy

- Instead of a simple "approve/reject" system, I recommend using the model to create a more sophisticated, tiered lending strategy.
- **Generate a 1-10 Risk Score:** Each applicant receives a score from our model.
- **Categorize into Risk Buckets:**
  - **Low-Risk (Scores 1-4):** Approve for standard loan terms.
  - **Medium-Risk (Scores 5-6):** Approve, but with adjusted terms (e.g., slightly higher interest rate, lower loan amount).
  - **High-Risk (Scores 7-8):** Route for manual review by a senior loan officer.
  - **Very High-Risk (Scores 9-10):** Reject.
- **The Business Benefit:** This approach **maximizes approvals** and revenue by not outright rejecting borderline cases. It protects the business by pricing risk appropriately and mitigates the impact of false positives.

Risk Score	Category
1 - 4	Low-Risk
5 - 6	Medium-Risk
7 - 8	High-Risk
9 - 10	Very High-Risk

Next Steps: Validation and Future Synergy

- **Validate with an A/B Test:** Conduct a live champion-challenger test to measure the model's real-world impact on default rates and profitability.
- **The Most Important Step - Combine Both Tasks:** The greatest opportunity for improvement lies in **combining the work from Part 1 and Part 2.** By enriching the application data with the powerful behavioural features engineered from the JSON credit reports, a single, unified model can be created that will be significantly more accurate and robust.
- **Monitor and Iterate:** Deploy the model and continuously monitor its performance, retraining as necessary.