PRESENTATION
CREDIT RISK ANALYSIS

Data Science Internship Project - Fikri Budianto
ID/X Partners X Rakamin Academy
Link Presentasi : https://youtu.be/9CZvTf7FFPM

INTRODUCTION

ABOUT

ID/X PARTNERS

IDX Partners, formally known as PT IDX Consulting, is an independent Indonesian consulting firm specializing in data analytics, decisioning, and regulatory technology (RegTech) solutions established in 2006.

The firm primarily assists financial services institutions, including top banks, multifinance companies, fintechs, and insurers, in leveraging advanced analytics and AI/ML-based solutions to optimize profitability and business processes.

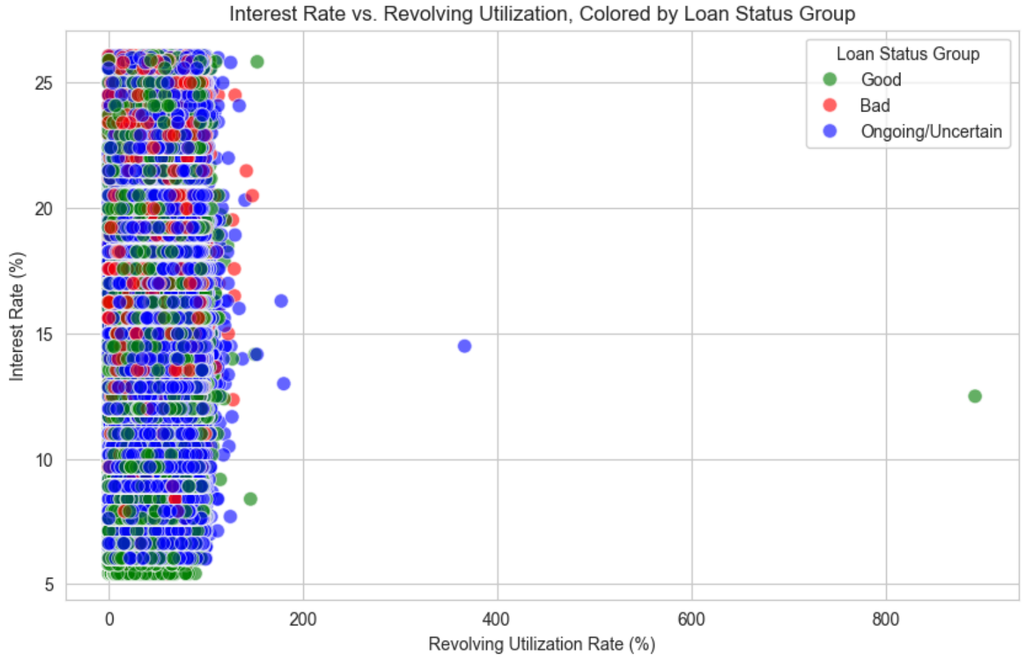
PROJECT CONTEXT

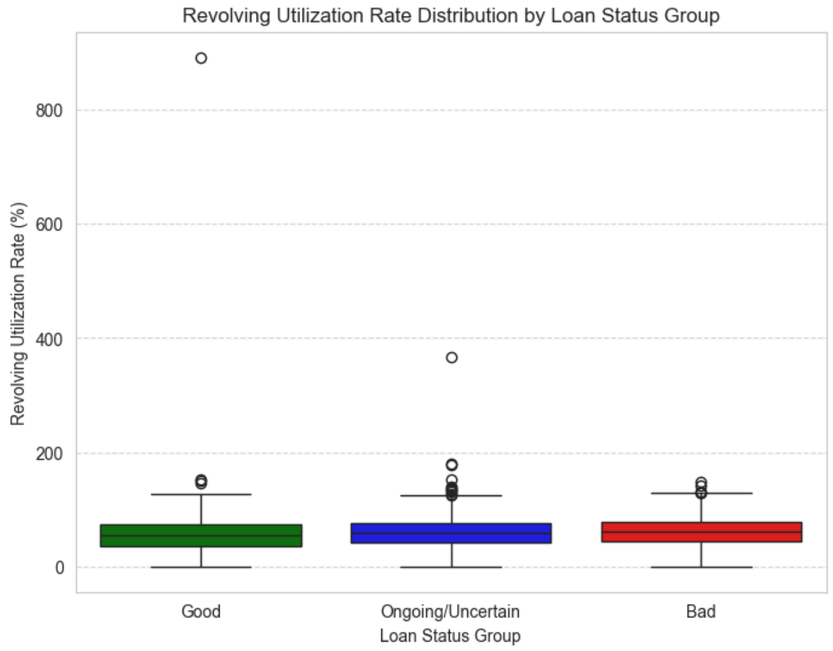
This project aims to analyze the credit risk of a borrower using a dataset provided by a lending company. The ultimate goal is to construct a robust predictive model that accurately forecasts the likelihood of a loan applicant defaulting (Charge Off/Bad Loan)

EXPLORATORY DATA ANALYSIS

INSIGHT 1

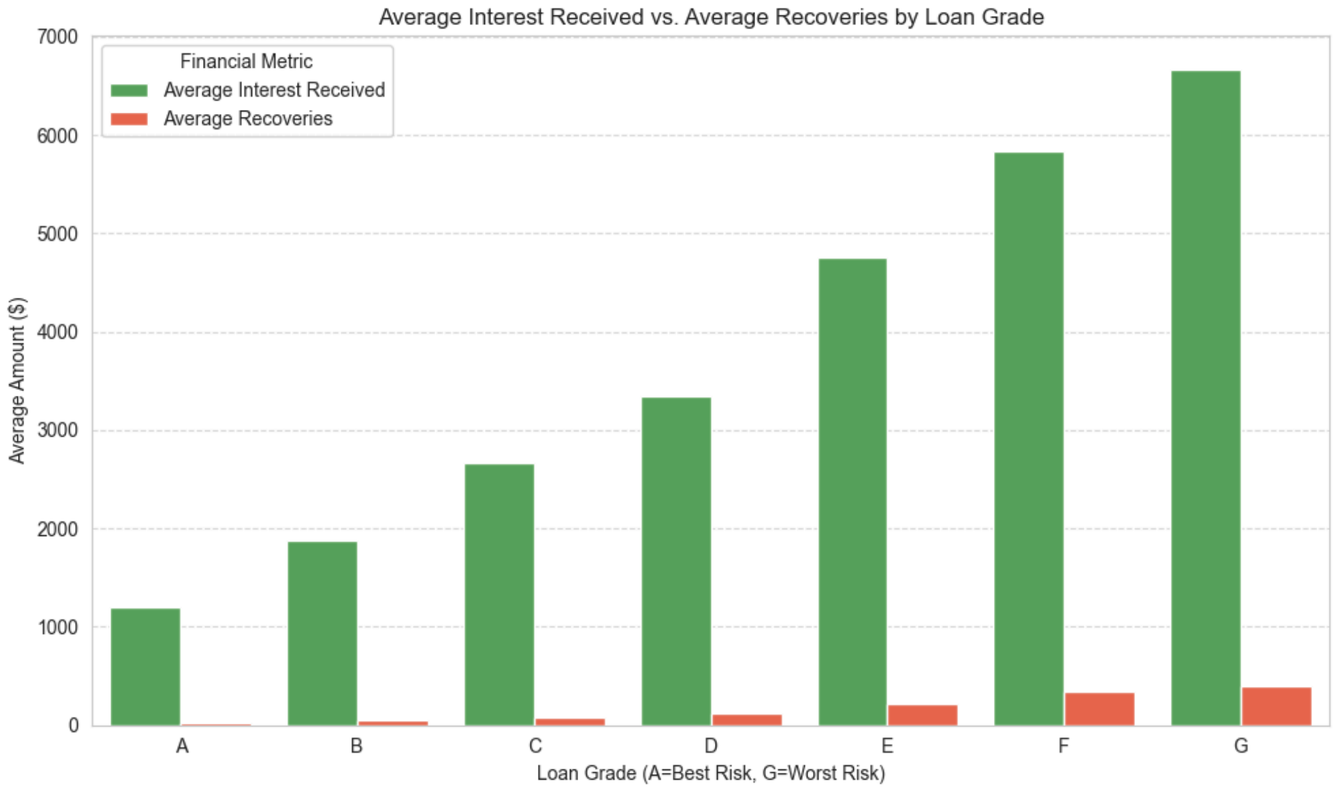
- The highest concentration of Bad Loans occurs in the high-risk segment, specifically where Revolving Utilization exceeds approximately 60% and Interest Rates are highest (above approximately 15%).
- The median utilization rate for defaulted loans is significantly higher than for healthy loans, confirming high debt usage is a primary financial strain indicator.
- The utilization distribution for 'Ongoing/Uncertain' loans closely aligns with 'Good' loans, suggesting the current book is not facing generalized distress from high credit usage.





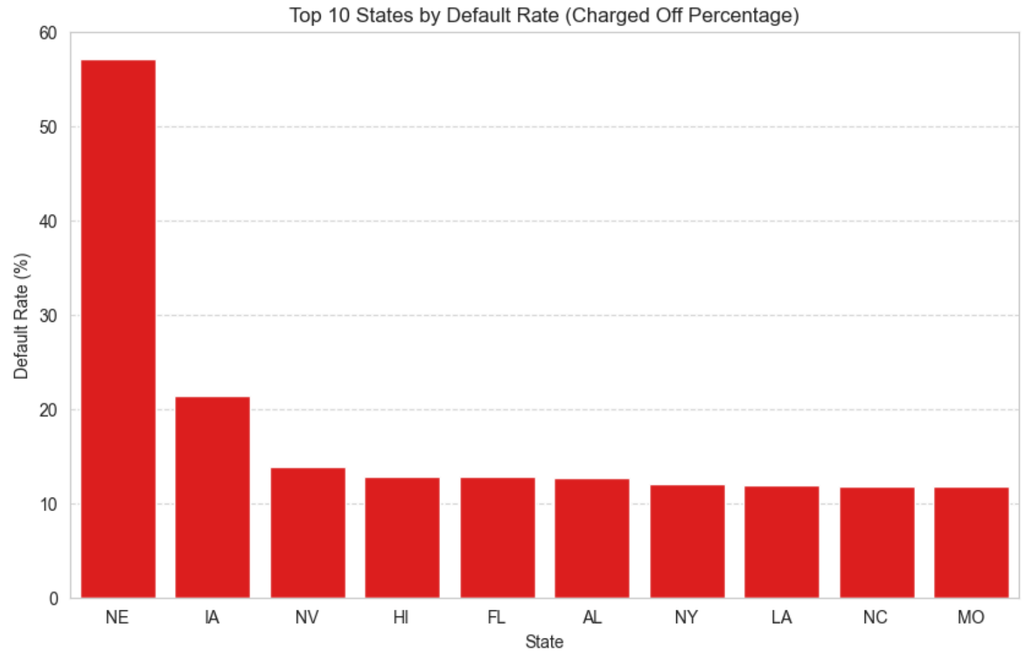
INSIGHT 2

- The highest net profitability is typically achieved in the medium-risk grades (B or C), as the moderately high interest charged outweighs the manageable level of losses.
- Recoveries (losses) escalate sharply from Grade D onward, with the highest-risk grades (F and G) requiring the largest recovery efforts, significantly eroding their potential revenue
- The lowest-risk grades (A and B) offer stability but yield the lowest revenue (total_rec_int), making them less attractive for maximizing portfolio returns.



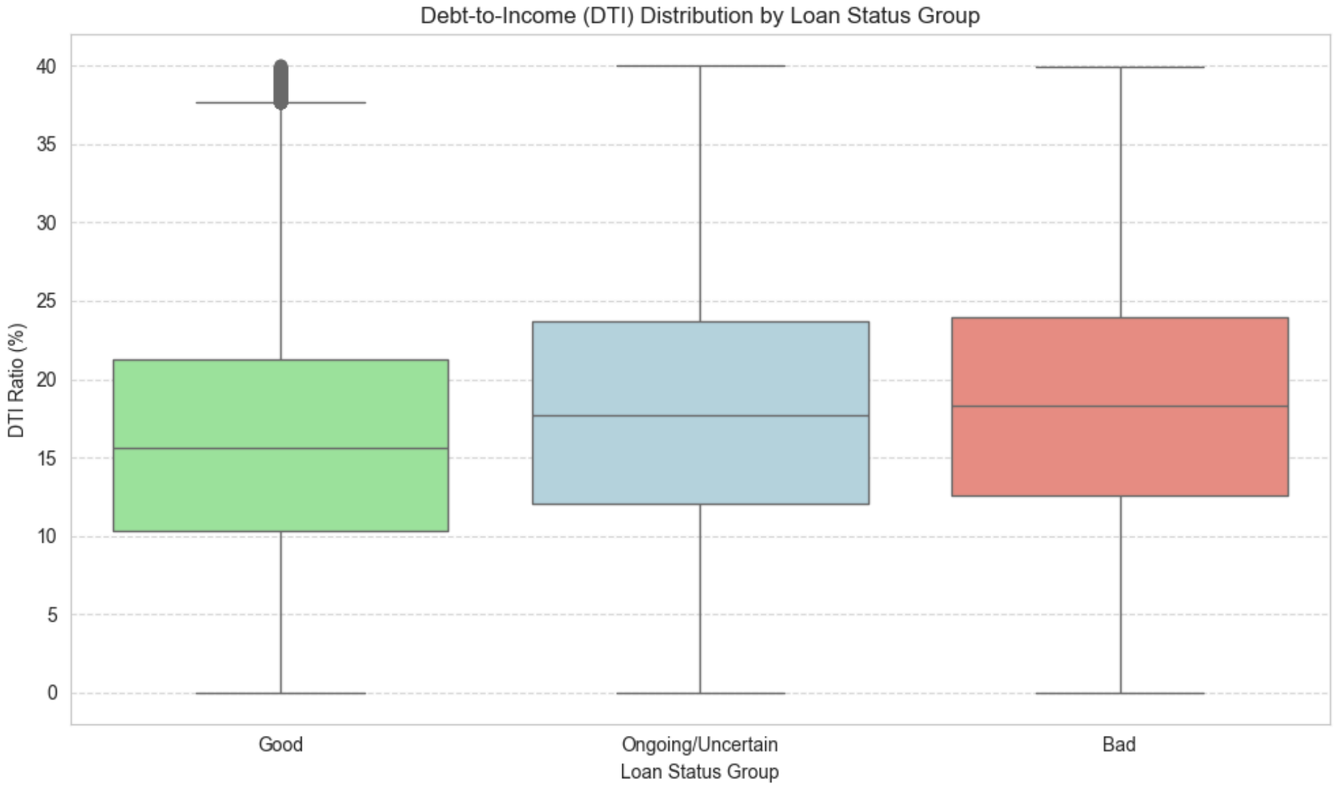
INSIGHT 3

- While California (CA) and Texas (TX) dominate loan volume, they do not appear in the top 10 states by Default Rate.
- States like Nevada (NV), Florida (FL), and Georgia (GA) exhibit a disproportionately high default rate, signaling regional economic or underwriting risks that are not adequately accounted for.
- Mitigation strategies (e.g., stricter underwriting or higher interest rates) must be applied to these high-risk states, even if their total loan volume is lower.



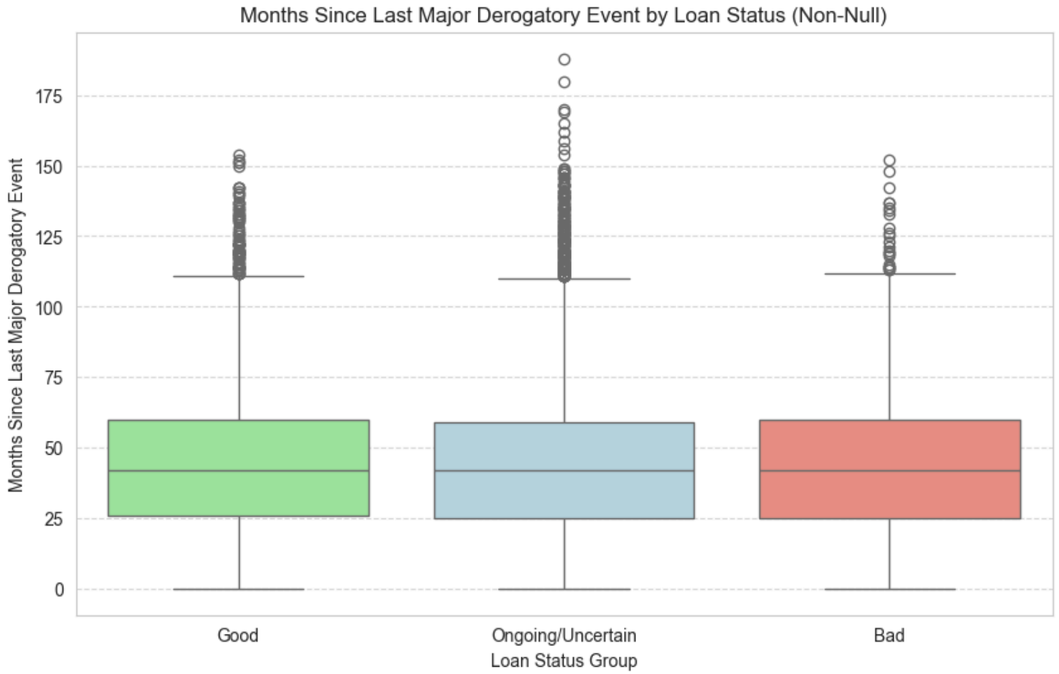
INSIGHT 4

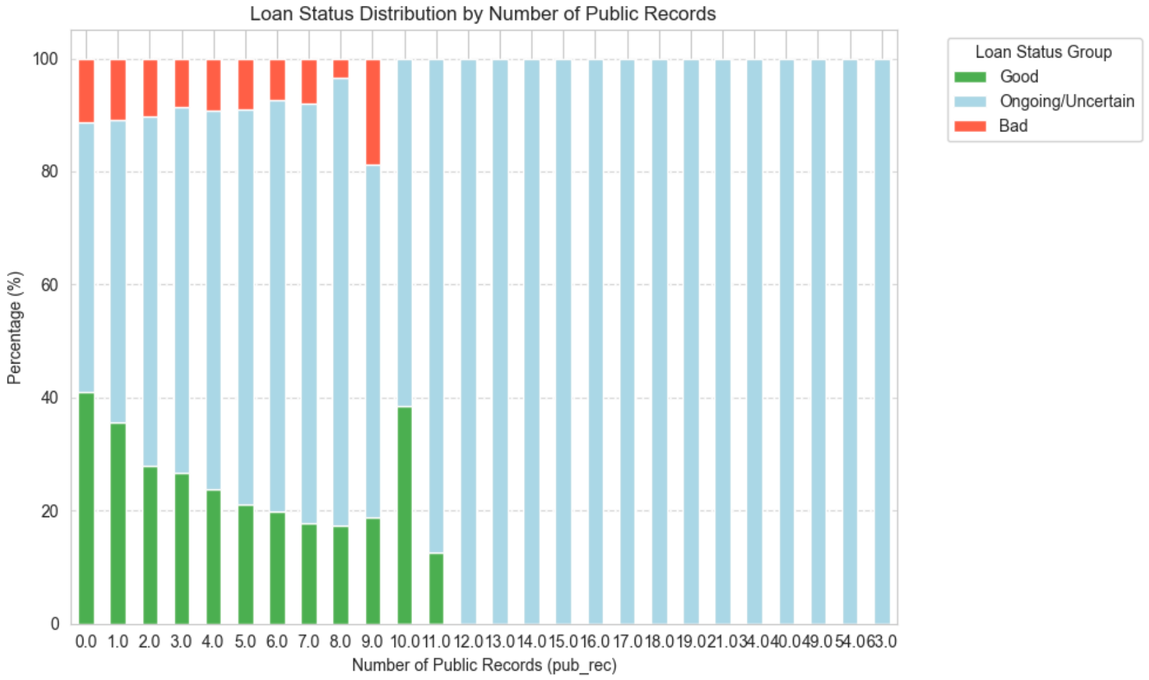
- The median Debt-to-Income (DTI) ratio for 'Bad' loans is significantly higher than for 'Good' loans, confirming that high debt burden is a major default predictor.
- The entire DTI distribution for the 'Bad' group is shifted upwards, indicating that borrowers with limited disposable income are consistently more prone to financial distress.
- The DTI of 'Ongoing/Uncertain' loans mirrors the 'Good' loans, suggesting DTI is a reliable metric for flagging potential issues early.



INSIGHT 5

- The median time elapsed since the last major derogatory event (mths_since_last_major_derog) is significantly lower for defaulted loans, proving that recent severe credit issues are a powerful precursor to failure.
- Loans with one or more public records (pub_rec) have a notably higher percentage of the 'Bad' category compared to those with zero records, confirming a history of public credit issues strongly predicts future default





DATA PREPROCESSING & MODELLING

INITIAL CLEANING, IMPUTATION, AND SCALING

- Handling Missing Values (Imputation): Missing values in Numerical columns were imputed using the median to minimize outlier bias. Missing values in Categorical columns were imputed using the mode (most frequent value).

- High Cardinality Cleanup: Highly sparse/textual columns and unique identifiers (e.g., url, emp_title, desc, title) were dropped due to low predictive value and high memory cost.

- Numerical Feature Scaling: All continuous numerical columns were scaled using StandardScaler (Z-score normalization), ensuring they all have a mean of 0 and a standard deviation of 1, which is necessary for model stability.

TARGET DEFINITION AND FEATURE ENCODING

- Target Transformation (Binary): The primary modeling objective was redefined from predicting the grade to a Binary Classification problem. A custom mapping was applied to the loan_status column, creating a new target y (where 0 = Good Loan and 1 = Bad/Troubled Loan).
- Feature Encoding: Categorical features were converted to numerical format using a strict cardinality threshold:
- Low Cardinality (<= 20 unique values): Encoded using One-Hot Encoding (OHE) (e.g., purpose, term).
- High Cardinality (>= 20 unique values): Encoded using Label Encoding (LE) (e.g., zip_code, addr_state, and all date columns), overwriting the original columns.

FEATURE SELECTION AND SPLITTING

- Leakage Removal: Columns that directly leak the target outcome, such as the original loan_status and payment history fields, were dropped/excluded from the feature set X.
- Multicollinearity Reduction: The final feature set X was checked for high inter-feature correlation (multicollinearity). 7 highly correlated features (e.g., member_id with id, funded_amnt with loan_amnt, and various payment/principal columns) were identified and removed using a threshold of 0.90.
- Train-Test Split: The processed data (X and y) was split into training and testing sets (X_train, X_test, etc.) using a stratified approach to maintain the proportion of "Bad Loans" in both sets.
- Feature selection: Only the features with importance > median will be kept. Median Feature Importance Threshold: 0.0013, Selected feature count: 41

MODELLING

- Binary Target & Imbalance: The problem was simplified to Binary Classification (Good vs. Bad Loan). We addressed the severe 7.4:1 class imbalance by using Class Weights in all models to prevent biased predictions.
- Hyperparameter Tuning: We efficiently optimized five strong models (XGBoost, CatBoost, AdaBoost, Logistic Regression, MLP) using RandomizedSearchCV and Stratified K-Fold cross-validation on the training data.
- Unbiased Performance: The final metrics (Accuracy, F1-Score) were calculated solely on the unseen Test Data, ensuring an unbiased assessment of the model's true predictive power on new loan applications.

MODEL PERFORMANCE

📊 Model Performance Comparison (after Hyperparameter Tuning):

	Model	Accuracy	Precision	Recall	F1 Score	\
0	XGBoost	0.9868	0.9867	0.9868	0.9866	
1	CatBoost	0.9790	0.9797	0.9790	0.9793	
2	MLP Classifier	0.9750	0.9747	0.9750	0.9749	
3	Logistic Regression	0.9218	0.9413	0.9218	0.9277	
4	AdaBoost	0.9346	0.9391	0.9346	0.9233	

RECOMMENDATIONS

RECOMMENDATIONS

- Implement the XGBoost/CatBoost model directly into the approval workflow to assign a precise risk probability to every new application.
- Strictly tighten underwriting rules around the strongest indicators: DTI and Revolving Utilization (revol_util) should not exceed the median levels observed in defaulted loans.
- Instantly flag or auto-reject applications originating from high-risk states (e.g., NV, FL, GA) and those showing recent major derogatory credit events.
- Develop a new, objective scorecard based on the model's Feature Importance to transparently justify all acceptance and rejection decisions.

THANK YOU