LOAN CREDIT RISK PREDICTION

Author: Bintang Phylosophie | Data Scientist Intern

Project Overview

Id/x partners is a private high growth organization headquartered in Jakarta and is a leading consulting firm for Data, Analytics & Decisioning solution in Asia-Pacific region. I am involved in a project from a lending company (multi finance), where the client wants to improve the accuracy of assessing and managing credit risk, so that they can optimize their business decisions and reduce potential losses. This can be done by developing a machine learning model that can predict credit risk based on the provided dataset, which includes data on approved and rejected loans.

Data and Business Understanding

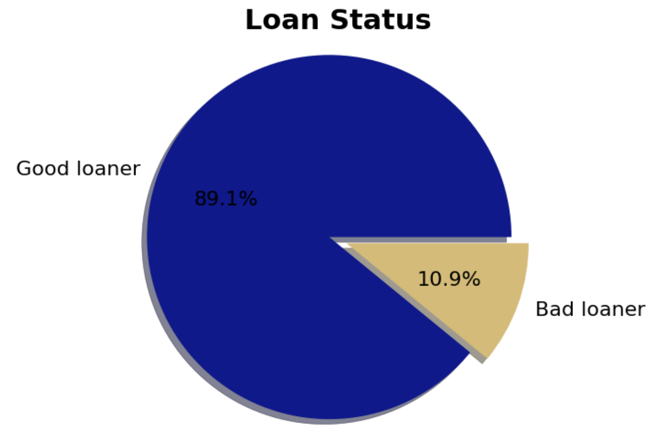
Dataset Information:
This dataset contains **466.285** loan information from a lending company namely [LendingClub] (https://www.lendingclub.com/) from 2007 to 2014. Contains 75 columns (float, integer, and object types), 5 columns contain date/time values

Attribute Information:

- **Identifier:**
  `id` and `member_id` is unique ID that each of which is an ID for loan listing and ID for the loaner member
- **Clients Attributes:**
  Client Information, Loan Purpose, Location, Loan Information
- **Target:**
  `loan_status` has several values, such as:
  - `Current` means current payments
  - `Charged Off` means the payment is in default so that it is written off
  - `Late` means late payment is made
  - `In Grace Period` means in grace period
  - `Fully Paid` means payment in full
  - `Default` means payment is stuck

  Later `loan_status` will be categorized as 'good loaner' and 'bad loaner'.

What Happened?

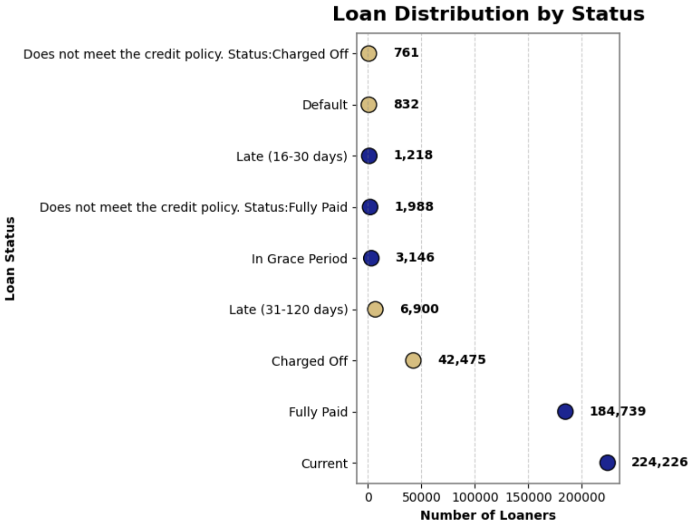


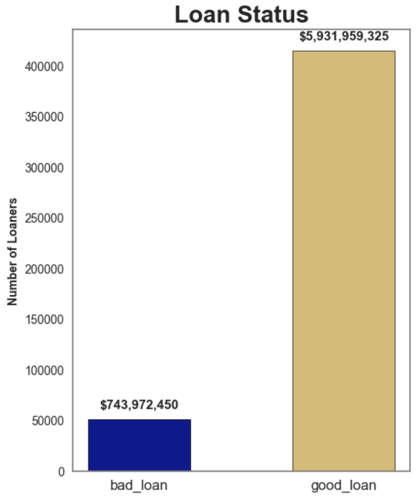
'Good loaners' is when the loan status is:

- **Current**, **fully paid**, **late < 30 days**,
- **In Grace Period**
- **does not meet the credit policy with status fully paid**.

Otherwise is 'Bad loaners' :

- **Charged off,**
- **Default**
- **Does not meet the credit policy. status: charged Off**
- **Late (31-120 days**).





Good Loan total **415.317**
Bad Loan total **50.968**

Problem Statement

- **Problems:**
  $ 743,9 million in defaults originating from **10,9% of borrowers**, serves as a critical warning to strengthen the credit risk mitigation strategies.
  Loan companies are faced with two major decisions that carry two types of risk regarding approval decisions:
  1. Approving loans to applicants who are unlikely to repay their loans resulting in financial losses for the company.
  2. Disapproving loans to applicants who are likely to repay the loan resulting in business losses.
- **Goals:**
  The goals to be achieved in this case are as follows:
  - Accepting applicants who demonstrate strong creditworthiness and are likely to be reliable borrowers
  - Declining applicants who present a high risk of default or exhibit characteristics of poor repayment behavior.

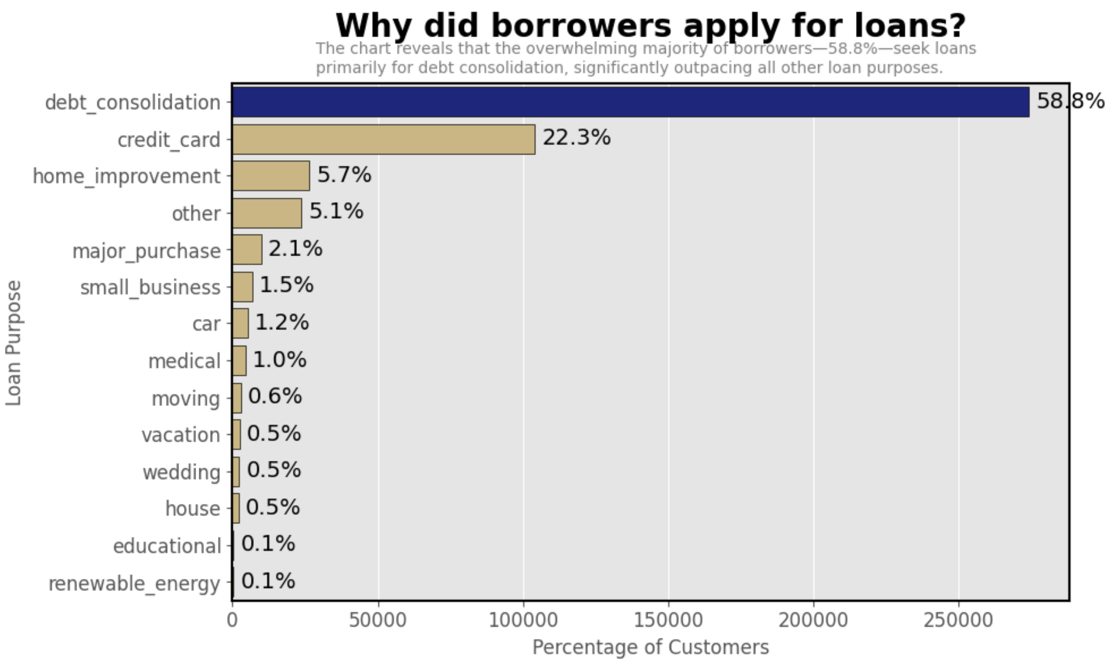
- **Objectives:**
  1. Predict whether the applicant is a good loaner or a bad loaner
  2. What makes the borrower indicated a bad loaner.
- **Metrics:**
  1. High-Risk Rate – Default Rate (DR)
  2. Total High-Risk Loans – Exposure at Default (EAD)
  3. Total Revenue – Expected Interest Income (EII
  4. Total High-Risk Loss – Expected Loss (EL)
  5. Net Revenue – Net Interest Income (NII)

Who is Applying for Loan?

Insight: In employment title, most of applicants have job in **Manager, Service,** and **Assistant**. Many applicant didn't write their employment title, so it's marked as **nan**.

Action: Tailor loan products and marketing strategies to better align with the needs and preferences of applicants roles with majority of applicants.

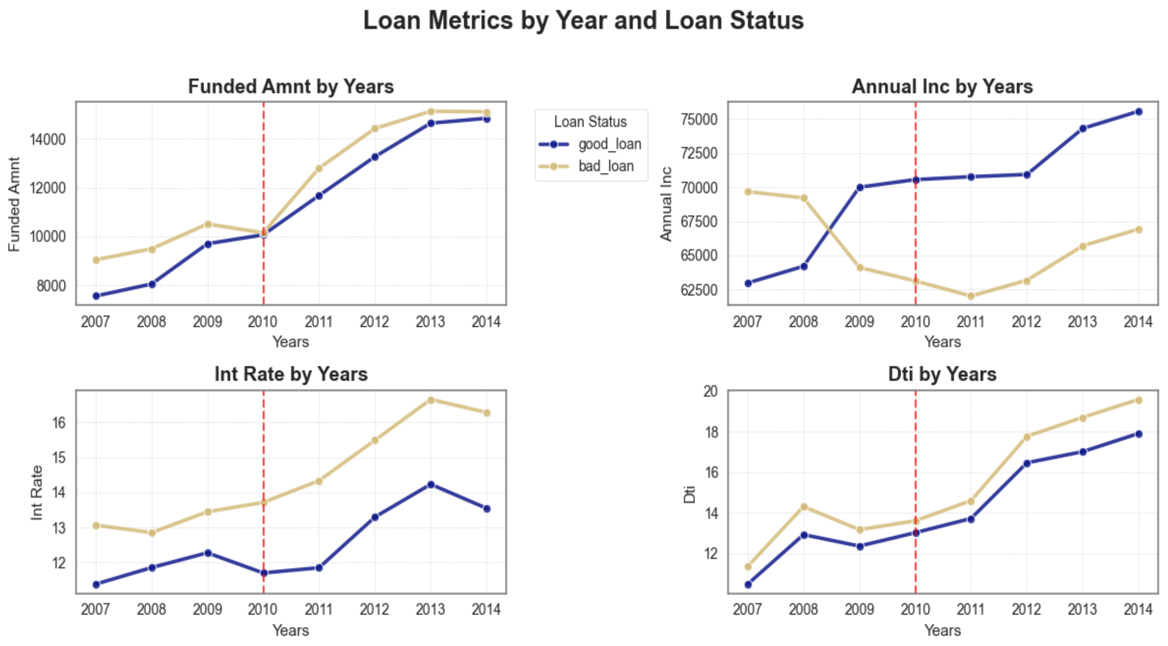
Why Did Borrowers Applying for Loan?



Insight: The chart reveals that the overwhelming **58.8%** of borrowers seek loans primarily **for debt consolidation**, significantly outpacing all other loan purposes.

Action: Develop targeted loan products and outreach campaigns focused on debt consolidation.

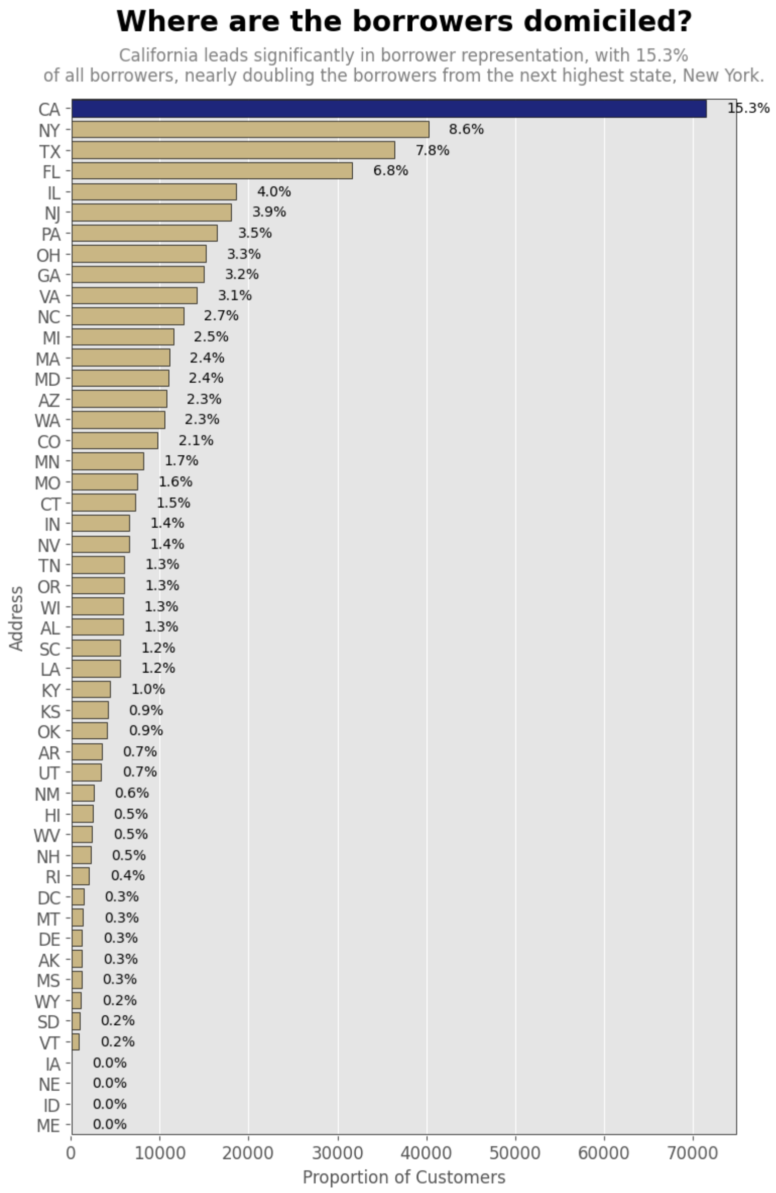
Loan Metrics by Year and Loan Status



The following graphs show the trends of key financial metrics from 2007 to 2014, separated by risk classification:

- Funded Amount **rose significantly** after 2010, especially for Good Loaners.
- Annual Income of Good Loaners **increased steadily**, while Bad Loaners remained flat.
- Interest Rates are **consistently higher** for Bad Loaners .
- Debt-to-Income Ratio shows a **persistent gap** between the two groups.

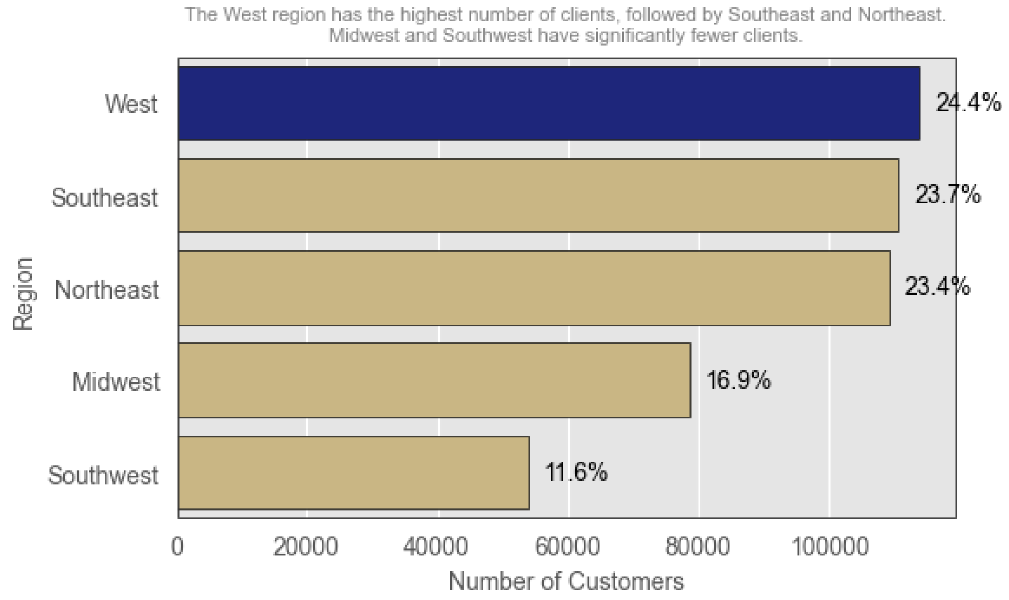
Where are the Borrowers Domiciled?



Insight: California dominates borrower representation, accounting for 15.3% of all borrowers — nearly double the next highest state, New York (8.6%). The top 5 states (CA, NY, TX, FL, IL) together represent over 40% of the total borrower base, showing a clear concentration in large, populous states. Meanwhile, many states (e.g., VT, SD, WY, AK) **contribute less than 0.5% each**, with some at effectively 0%.

Action: Focus marketing and loan efforts on top states like California, New York, and Texas, while expanding outreach in underrepresented states to reduce concentration risk.

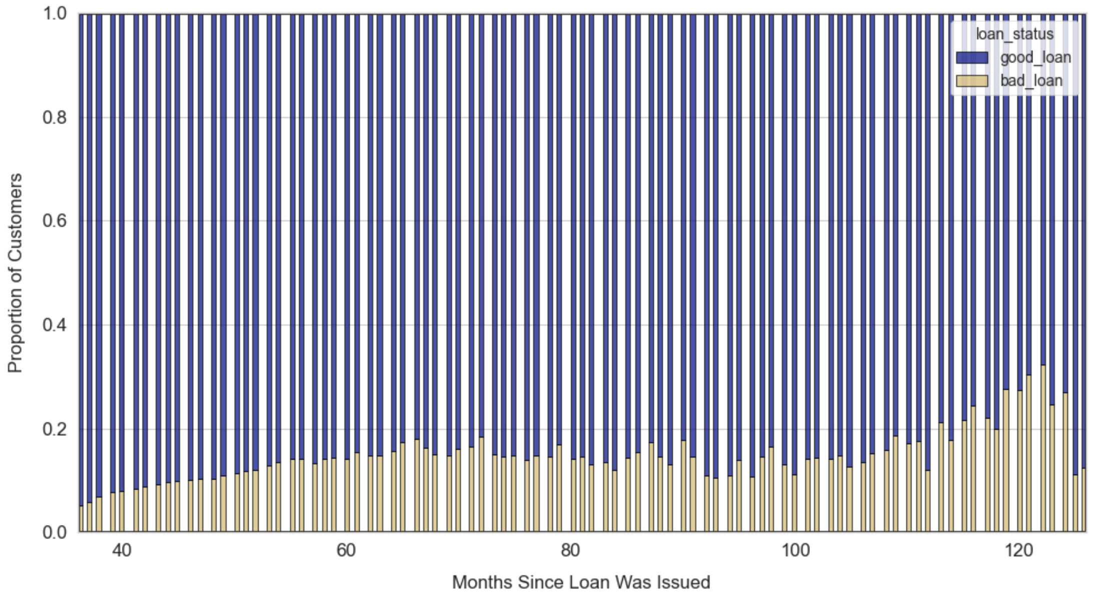
Borrower Distribution by Region



Insight: The **West region** has the highest number of clients (24.4%), followed by Southeast (23.7%) and Northeast (23.4%). Midwest and Southwest have significantly fewer clients.

Action: Marketing initiatives can target underrepresented regions like Southwest, which show low risk but untapped client potential.

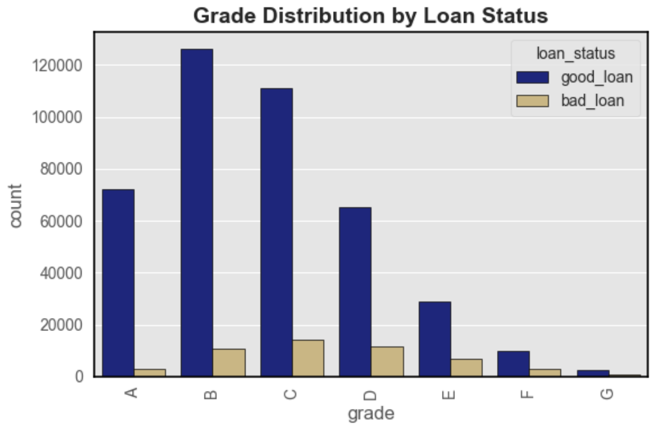
The Length of Time Since Loan Was Funded



Key Insight: As the loan age increases (months since issued), the proportion of bad loans gradually rises — especially after the 100-month mark, where bad loans visibly **increase to 25–30%** of total loans. Early months (30–60) show a much lower default rate, often **below 10%.**

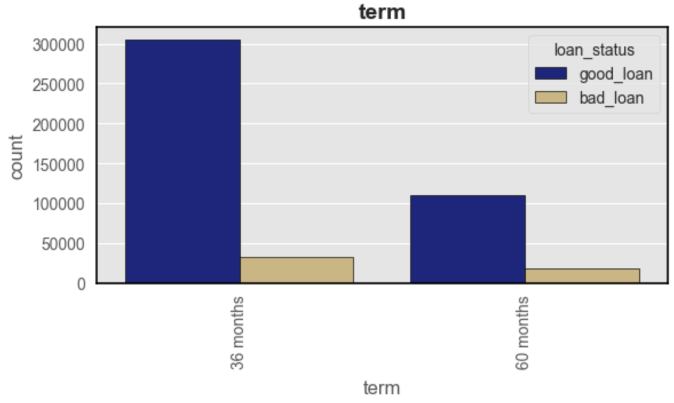
Action: Strengthen monitoring and risk management for older loans (especially >100 months) to proactively handle increasing default risk.

Bivariate Analysis



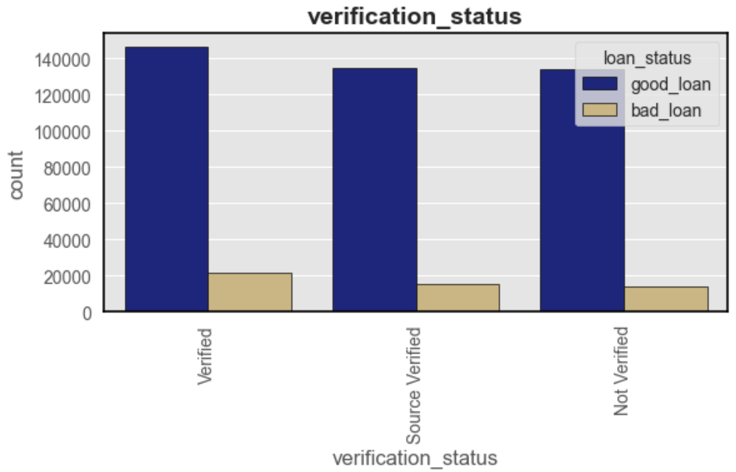
Insight: Grades B and C have the highest number of loans, both good and bad, but lower-grade loans (E, F, G) have a higher proportion of bad loans relative to good ones.

Action: Tighten approval criteria or increase interest rates for lower-grade loans (D–G) to offset higher default risk.



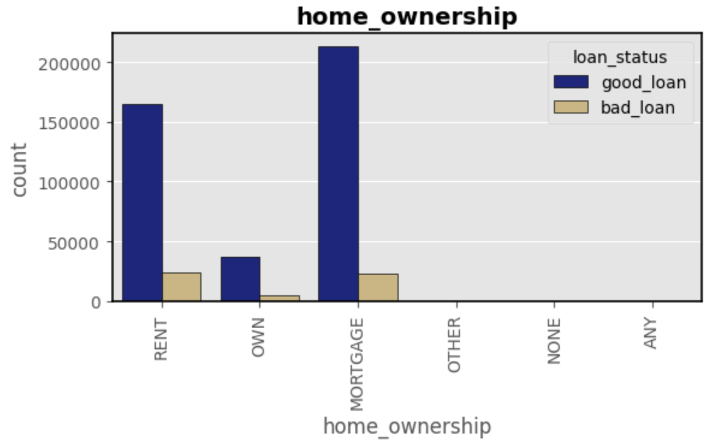
Insight: Loans with **60-month terms** show a higher proportion of defaults compared to 36-month loans.

Action: Prioritize offering 36-month terms to reduce default risk, or enhance risk pricing for 60-month loans.



Insight: Loans with **verified income** have slightly fewer defaults compared to unverified ones.

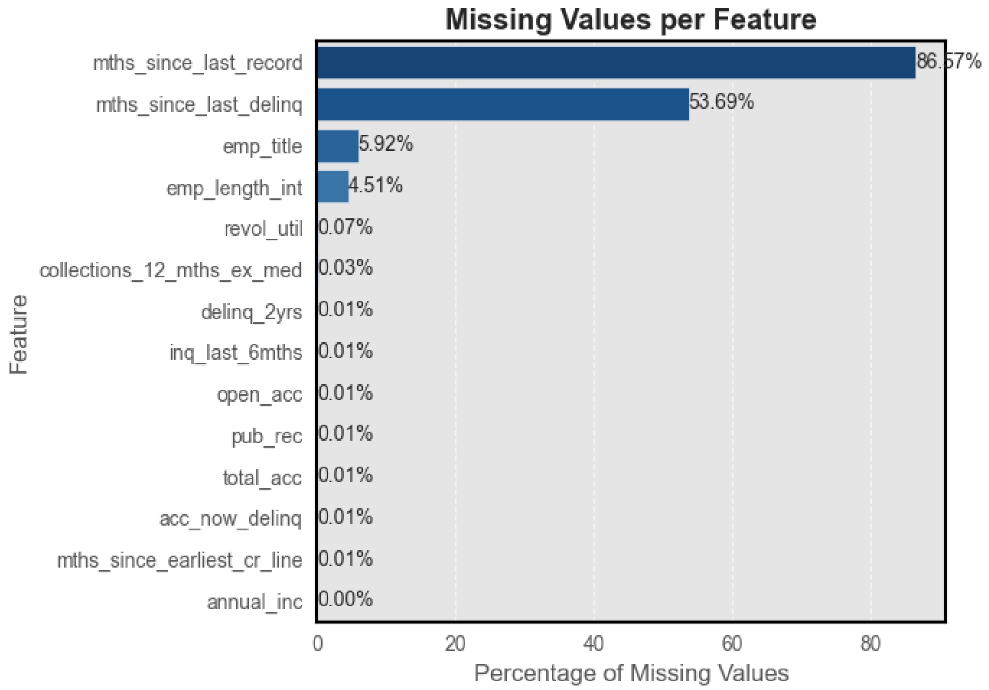
Action: Make income verification a standard requirement to improve loan quality and reduce risk.



Insight: Most borrowers rent or have **mortgages**. The bad loan rate is relatively higher among renters.

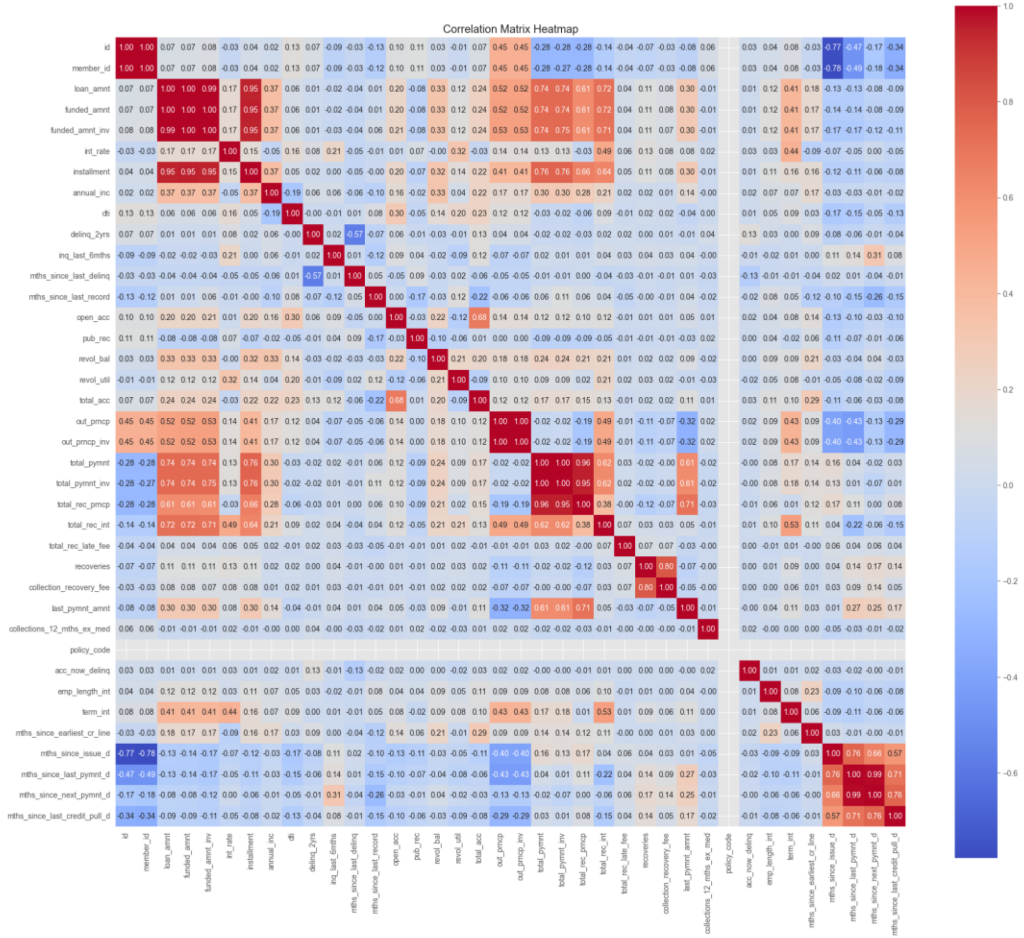
Action: Factor in homeownership status in risk assessment—consider tighter screening for renters or adjusting lending limits.

Data Preprocessing



- Several **unnecessary features have been removed**, including those representing unique identifiers, free-text fields, columns containing all null values, and others excluded based on expert judgment.
- Features with **very high or single cardinality were also excluded**.
- **Feature engineering** was applied to derive new variables from existing ones, including the target variable extracted from the 'loan_status' column; 1**= 'good_loan, 0 = 'bad_loan'.**
- The data types for datetime fields have been **properly adjusted**.
- The dataset contains **no duplicate** records.
- **Missing values have been imputed** based on appropriate strategies, using either the median or mode depending on the nature of the feature.

Feature Selection



- To prevent redundancy and potential multicollinearity, **highly correlated features were excluded** from the modeling process.
- Categorical **variables were encoded** using label encoding with the LabelEncoder() method.
- **Class imbalance in the target** variable was addressed using the SMOTE (Synthetic Minority Over-sampling Technique) algorithm.

- After splitting the dataset into **training and testing** sets with an 80/20 ratio, feature standardization was performed using **StandardScaler().**
- After completing all the preprocessing steps, the dataset with **28 features selected** is now ready for machine learning model development using several algorithms.

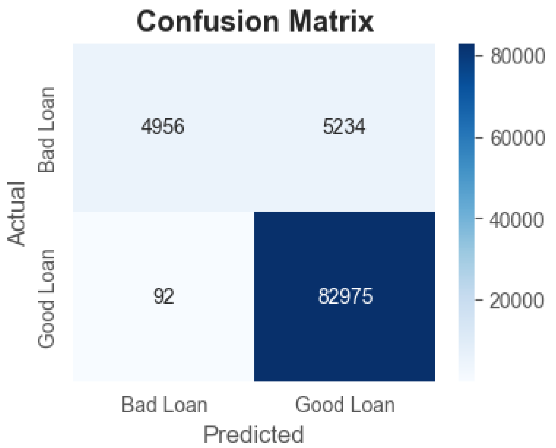
Model Development

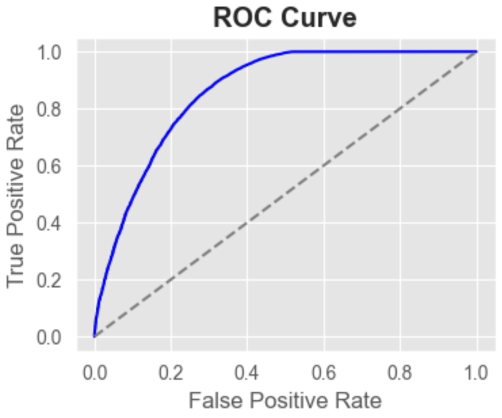
I use Random Forest, Logistic Regression, Decision Tree, XGBoost, Gradient Boosting for mode development.

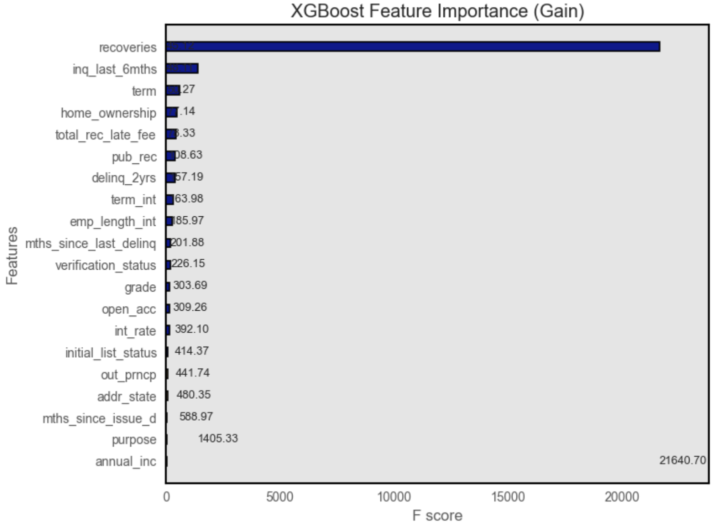
Model Evaluation

- To balance the trade-off between false positives and false negatives, the **F1 score** will be used as the primary evaluation metric, as it provides a more comprehensive measure of model performance in imbalanced classification tasks.
- However, **accuracy** will also be considered, as it offers an easily interpretable measure of overall model correctness. And **recall** to avoid approval of risky borrowers,
- increase revenue.
- In credit risk modelling, test performance is calculated using **the AUC metrics**.

Model	Accuracy	F1 Score	Recall	ROC AUC
Random Forest	0.942	0.968	0.995	0.872
Logistic Regression	0.902	0.944	0.945	0.810
Decision Tree	0.886	0.927	0.935	0.751
XGBoost	**0.944**	**0.968**	**0.969**	**0.902**
Gradient Boosting	0.942	0.952	0.968	0.860

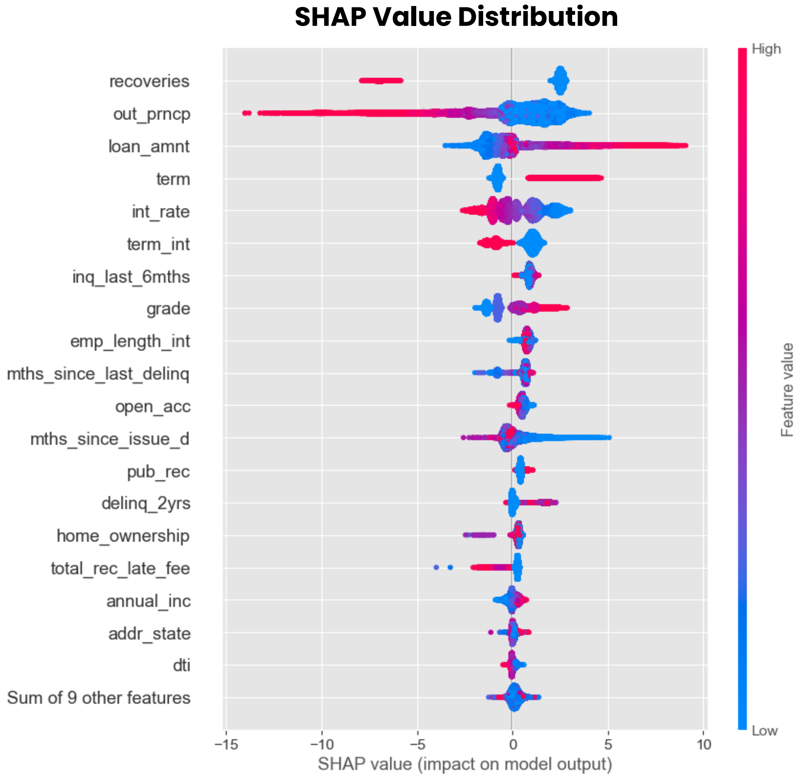






- True Negatives **(TN)** - correctly predicted bad loans: 82975
- False Positives **(FP)** - predicted bad, actually good: 92
- False Negatives **(FN)** - predicted good, actually bad: 5234
- True Positives **(TP)** - correctly predicted good loans: 4956

XGBoost gives the best performance.
Highest AUC **(90,2%)**, Strong generalization
High Recall **(96,9%)**, Effective high-risk detection
High F1-Score **(96,8)**, Balanced and reliable
Stable in both train & test, Low overfitting risk

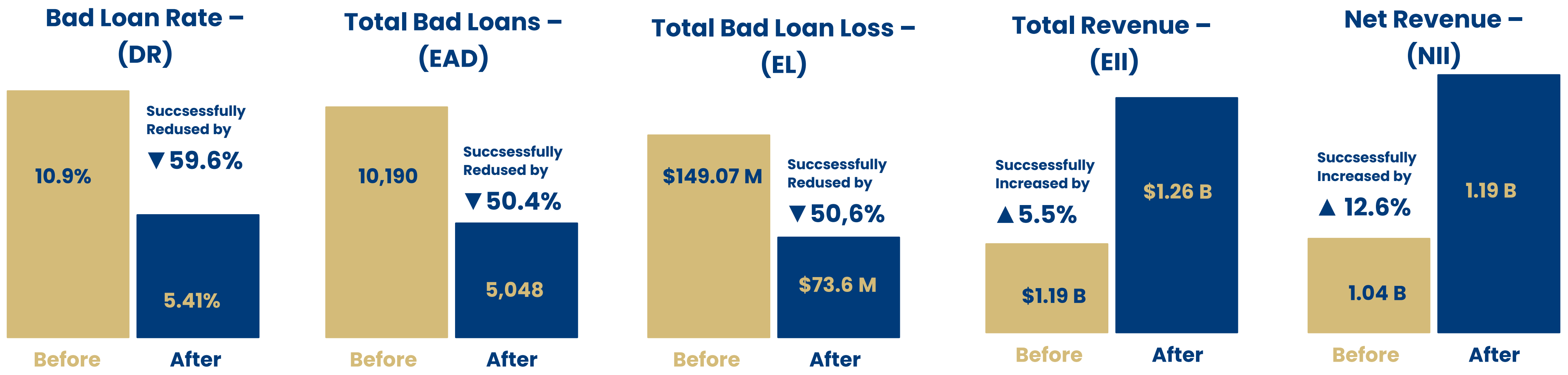


Features Importance & SHAP Value Distribution

- Most influential features: 'recoveries' (Indicates if a payment plan has been put in place for the loan), 'inq_last_6mths' (the number of inquiries in past 6 months (excluding auto and mortgage inquiries)), and 'term' (the number of payments on the loan. Values are in months and can be either 36 or 60.)
- SHAP (SHapley Additive exPlanations) explains how each feature shifts the prediction.
- High feature values (red) may reduce risk (positive SHAP), while low values (blue) may increase risk (negative SHAP).

Metrics Impact Simulation

TP = 4956 FP = 92 FN = 5234 TN = 82975 Total Clients = 466.140 Total Clients X: 93.257 Total Good Loan Clients (Before): 83067 Total Bad Loan Clients (Before): 10.190 Total Loan Amount: $149.069.950 Avg loan Amount: $14.313	**Bad Loan Rate - Default Rate (DR)** Bad loan rate before = total high risk before / total clients Bad loan rate after = (FP + TN) / total clients X	
	Total Bad Loans – Exposure at Default (EAD) EAD = total bad loan clients × avg loan	**Total Bad Loan Loss – Expected Loss (EL)** EL = expected bad loan clients × avg loan
	Total Revenue – Expected Interest Income (EII) EII = good loan clients × Avg loan	**Net Revenue – Net Interest Income (NII)** NII = EII – EL



Business Recommendation

Recoveries-Driven Credit Risk Model = Profit Engine

Insight:

'recoveries' is the most important feature in the XGBoost model, indicating past recovery performance is a strong predictor of loan success.

Strategy:

- Prioritize lending to applicants with strong historical recoveries
- Develop recovery-based scoring tier in underwriting
- Incentivize early settlements through rewards

Real Impact:

- Default rate reduced from **10.9% → 5.41%** (▼ 59.6%)
- Net Interest Income (NII) increased from **$1.04B → $1.19B** (▲ 12.6%)
- **$75M+ expected** loss prevented by focusing on recovery signals

Smart Income & Inquiry-Based Filtering = Churn Blocker

Insight:

Features like 'annual_inc', 'inq_last_6mths', and 'term' help identify risky segments—short-term loans, low income, and recent inquiries signal higher default potential.

Strategy:

- Real-time approval filters for risky applications
- Tiered interest rates based on inquiry recency and income
- Auto-decline thresholds for flagged combinations

Real Impact:

- Bad loans cut from **10,190** → **5,048** (▼ 50.4%)
- Expected Loss reduced by **$75.4M**
- Risk exposure cut in half while maintaining growth

Speed + Loyalty = Revenue Multiplier

Insight:

Reliable payment behavior (e.g. 'last_pymnt_amnt', 'total_pymnt', 'payment_time') correlates with high-value, low-risk customers.

Strategy:

- Instant approval for "loyal" profiles
- Cashback & reduced rates for consistent payers
- Live loyalty score to promote financial discipline

Real Impact:

- Revenue up +12.26% through loyalty targeting
- Onboarding speed improved
- Drop in high-risk client share

Conclusion

Data-Driven Insights:

- SHAP analysis and model simulations revealed **'recoveries'** and **payment behavior** as leading predictors of loan success, with **income and inquiry history** also playing key roles in identifying risk.
- Top features driving outcomes: 'recoveries', 'last_pymnt_amnt', 'total_pymnt', 'payment_time', 'annual_inc', 'inq_last_6mths', 'term'.

Three-Pronged Strategy:

1. **Recoveries-Based Risk Modeling:**
   Prioritize applicants with strong historical recoveries to lower default likelihood and prevent expected losses (▼$75M+).
2. **Smart Income & Inquiry Filtering:**
   Filter out high-risk profiles in real-time using income and inquiry data—cutting bad loans in half (▼50.4%).
3. **Loyalty-Driven Acceleration:**
   Use loyalty indicators for instant approvals, reduced rates, and higher retention—enabling faster onboarding and revenue growth.

Operational Enhancements:

- AI-powered segmentation supports **faster, more precise decisions**.
- Recovery-tier scoring and real-time flags improve risk detection.
- Transparent, data-backed policies boost trust and growth.

Main Impact & Business Value

- **Bad Loan Rate (DR)**: Reduced by **59.6%** (from 10.9% → 5.41%)
- **Total Bad Loans**: Cut by **50.4%** (from 10,190 → 5,048 clients)
- **Expected Loss (EL)**: Reduced by **$75.4M+**
- **Net Interest Income (NII)**: Increased **12.6%** ($1.04B → $1.19B)
- **Total Revenue (EII)**: Increased **5.5%** ($1.19B → $1.26B)
- **Customer Quality**: Drop in high-risk share, rise in loyal, low-risk clients

Click Here to See My Code

Linked In | GitHub | Portfolio