LOAN CREDIT RISK PREDICTION

Author: Bintang Phylosophie

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

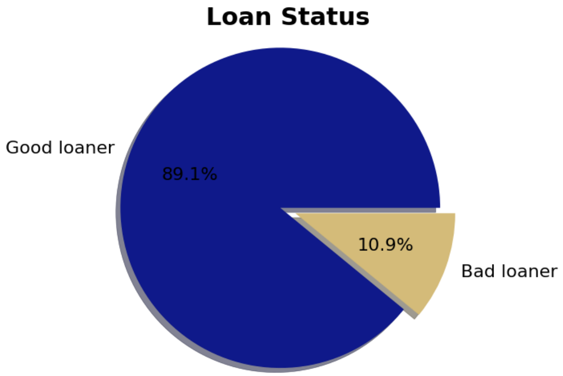
Dataset Information:
This dataset contains loan information from a lending company namely [LendingClub] (https://www.lendingclub.com/) from 2007 to 2014.
Aattribute Information:

- **Identifier:**
  `id` and `member_id` is unique ID that each of which is an ID for loan listing and ID for the loaner member
- **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'.

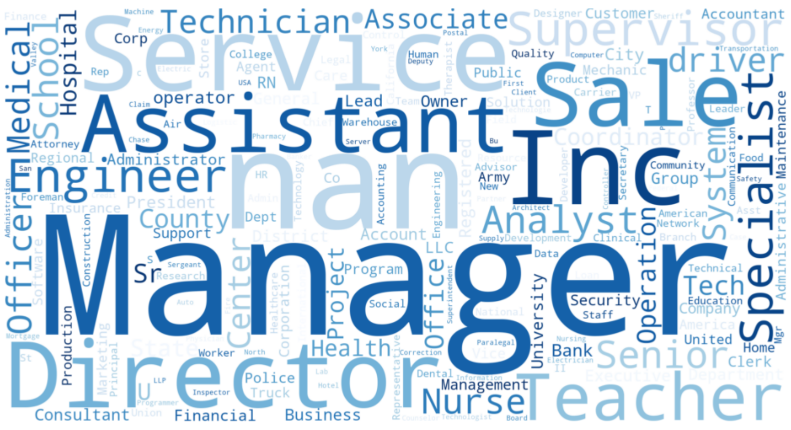
- **Goals:**
  The goals to be achieved in this case are as follows:
  - Accepting applicants who demonstrate strong creditworthiness and are likely to be reliable borrowers
  - Declining applicants who present a high risk of default or exhibit characteristics of poor repayment behavior.
- **Problems:**
  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.
- **Objectives:**
  1. Predict whether the applicant is a good loaner or a bad loaner
  2. What makes the borrower indicated a bad loaner

What Happened?



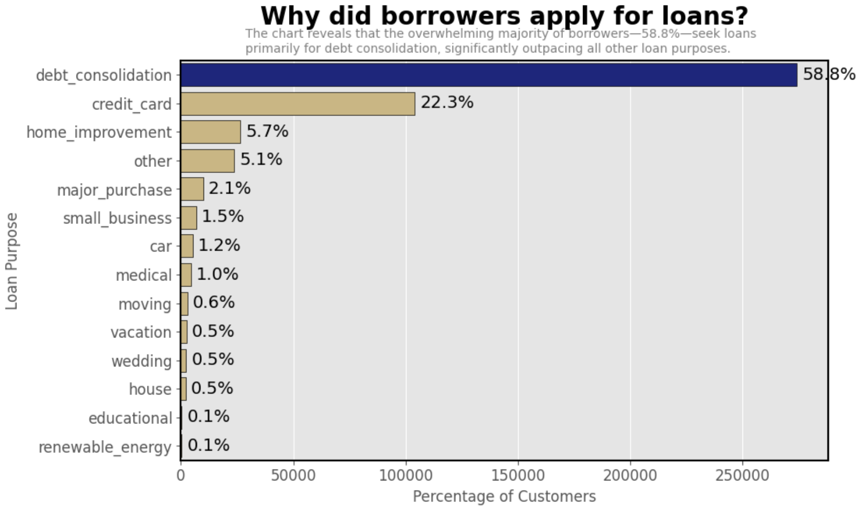
'Good loaners' is when the loan status is **current**, **fully paid**, **late < 30 days**, **In Grace Period & does not meet the credit policy with status fully paid**. Otherwise is 'Bad loaners' (such **charged off, default, does not meet the credit policy. status: charged Off, late (31-120 days)**).

Who is Applying for Loan?



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**.

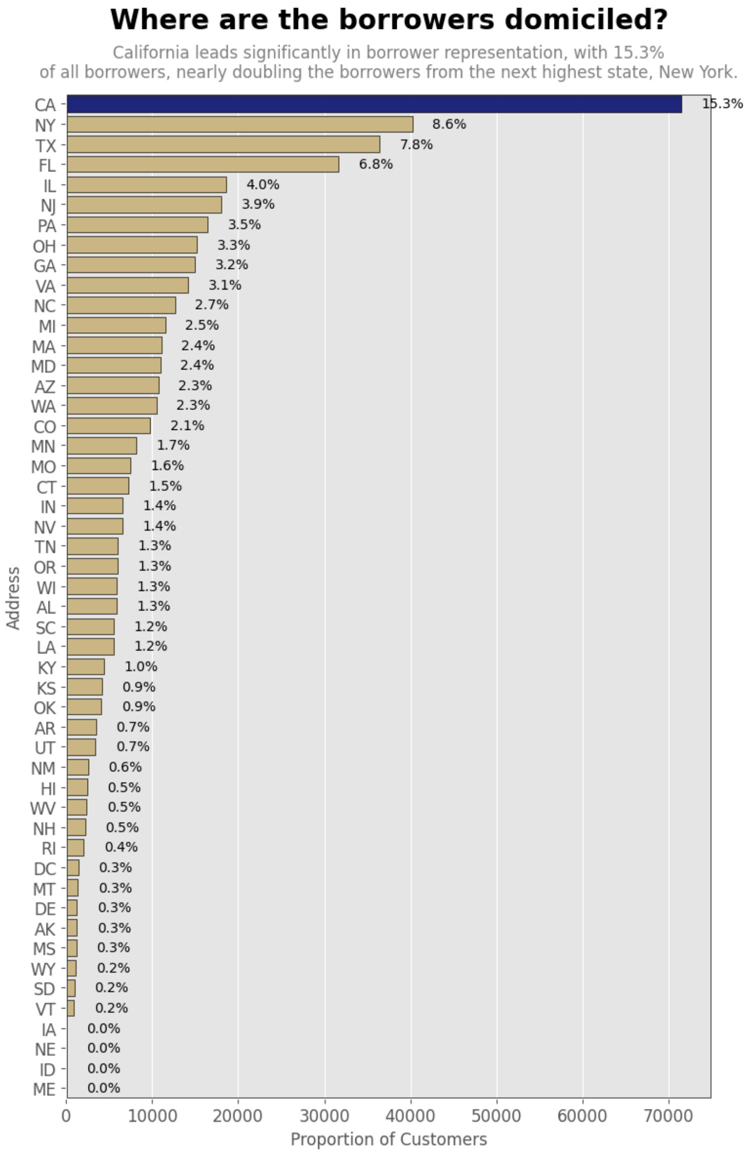
Why Do Borrowers Applying for Loan?


Why did borrowers apply for loans?

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

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

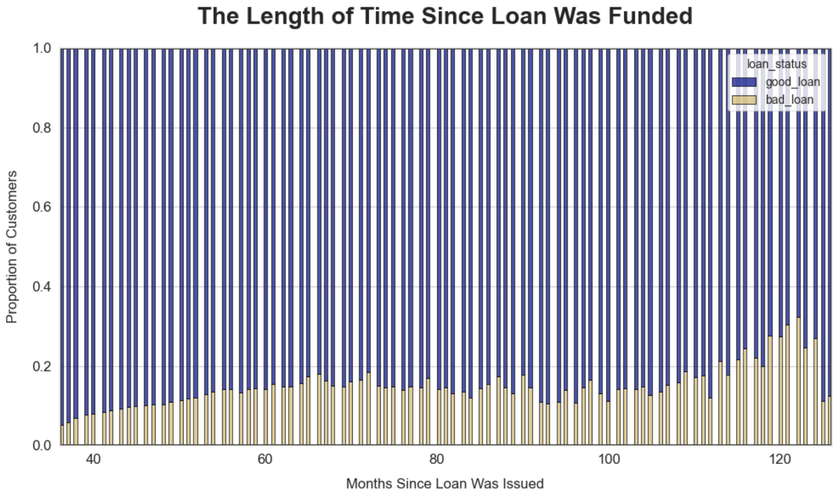
Where are the Borrowers Domicilied?


Where are the borrowers domiciled?

California leads significantly in borrower representation, with 15.3% of all borrowers, nearly doubling the borrowers from the next highest state, New York.

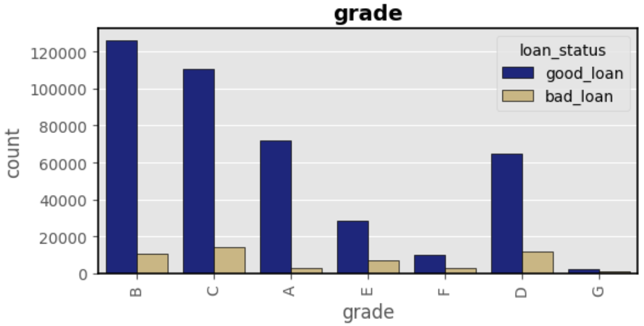
The chart shows that most borrowers come from a few key states, with **California leading by a large margin at 15.3%.** This is almost twice as many as **New York (8.6%),** followed by **Texas (7.8%)** and **Florida (6.8%).** These four states make up a big portion of all borrowers. Most other states have much smaller shares, with some states like Iowa, **Nebraska, Idaho, and Maine having almost none**. This suggests that borrowers are mostly from large, heavily populated states.

The Length of Time Since Loan Was Funded

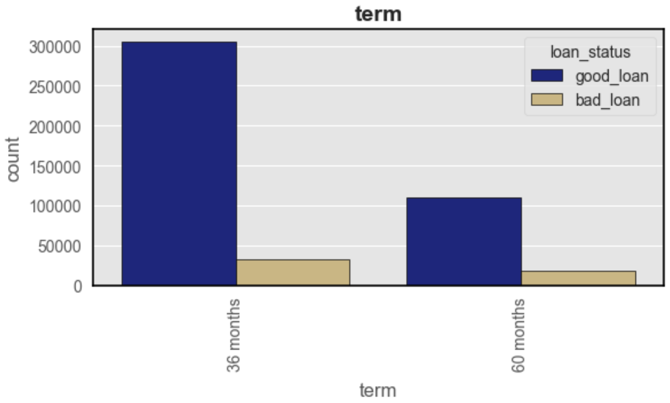


The chart shows that in earlier months, **nearly all loans were classified as good loans**, but over time, the proportion of **bad loans has steadily increased**. More recent loans (closer to the present) show a notable rise in defaults, with bad loans making up over 20–30% of issued loans in the later months.

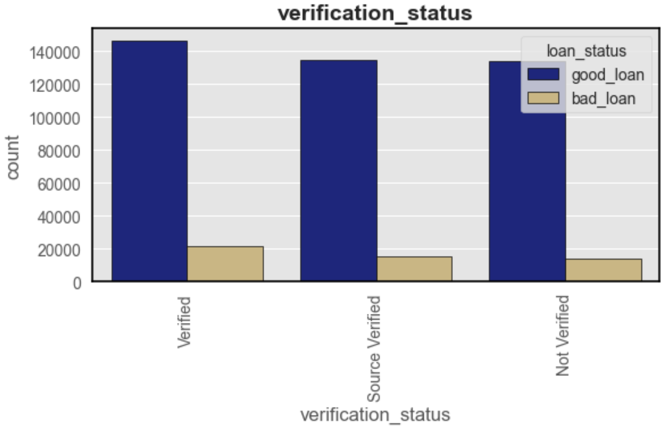
Bivariate Analysis



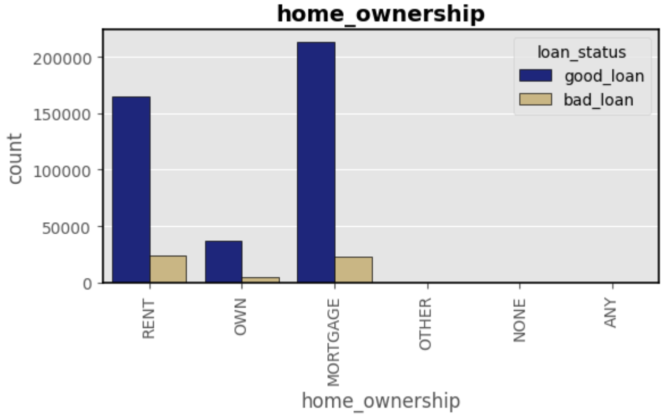
Lower credit grades (E, F, G) show a higher proportion of **bad loans**, while grades A to C, especially B, are dominated by **good loans**.



Loan risk with a span of **60 months** as a **high level** of adverse impact.

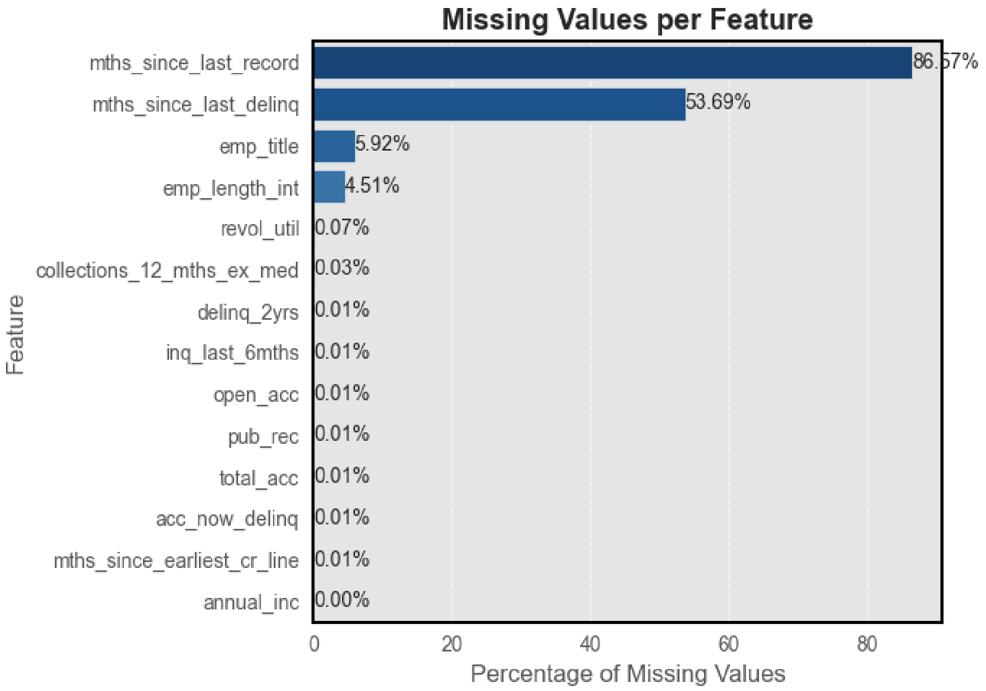


Loans with **verified income have fewer defaults,** while those without verification show a noticeably higher rate of bad loans.



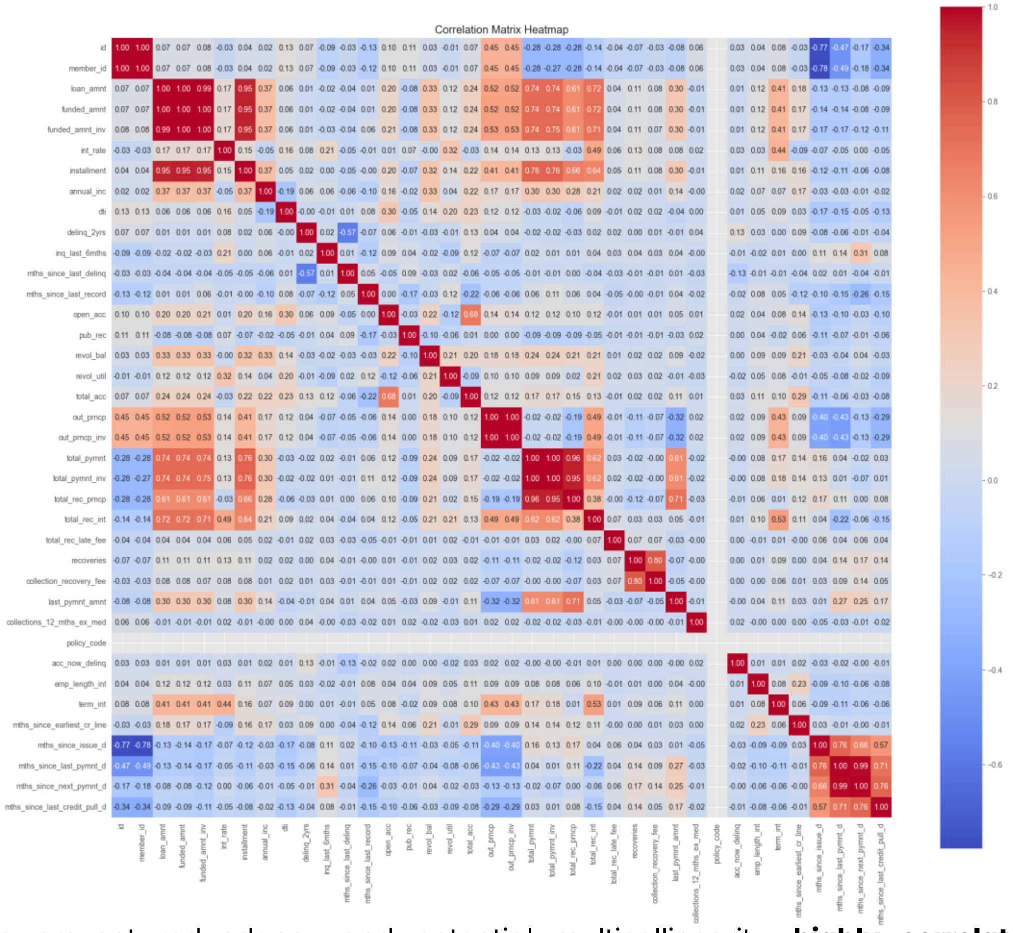
Bad loans are more common among renters, whereas borrowers with a mortgage or **who own a home tend to have better loan** performance.

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.
- 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 is now ready for machine learning model development using several algorithms.

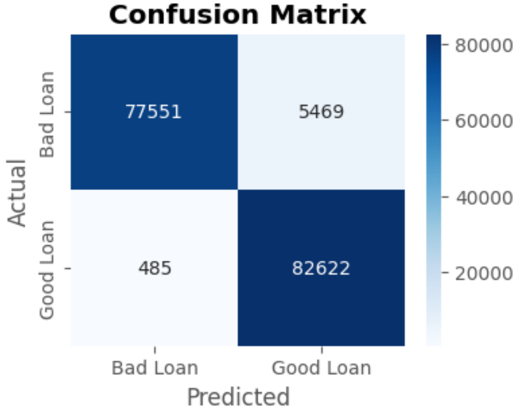
Model Development

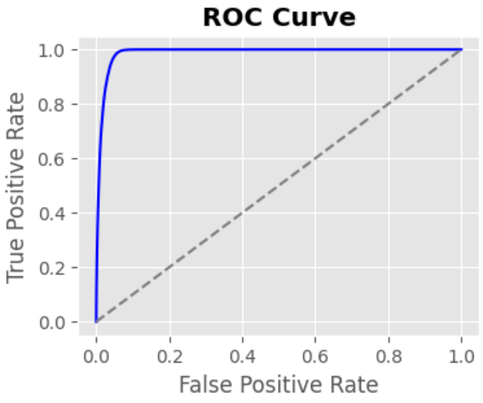
I use Random Forest, Logistic Regression, Decision Tree, XGBoost, Gradient Boosting for mode development.

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.
- In credit risk modelling, test performance is calculated using **the AUC metrics**.

Model	Accuracy	F1 Score
Random Forest	0.964	0.965
Logistic Regression	0.822	0.842
Decision Tree	0.928	0.927
XGBoost	**0.967**	**0.968**
Gradient Boosting	0.950	0.952





XGBoost gives the best performance. And the ROC Curve performance reached 0.988 using XGBoost.

Conclusion

- After evaluating multiple classification models, I have chosen **XGBoost** as the final model. It outperformed the others in terms of accuracy, ROC AUC, and precision, making it the most reliable choice for this case.
- Additionally, XGBoost achieved an impressive **F1 score of 0.97**, indicating a strong balance between precision and recall. These results demonstrate its effectiveness in handling the classification problem with both high predictive power and robustness.
- It also achieved the fastest execution time among all models, **completing in just 23.89 seconds** compared to the average runtime of 8 minutes.

Click Here to See My Code